BMJ Open Sport & Exercise Medicine

Visual assessment of movement quality in the single leg squat test: a review and meta-analysis of inter-rater and intrarater reliability

John Ressman,[1] Wilhelmus Johannes Andreas Grooten,[1,2] Eva Rasmussen Barr[1]

To cite: Ressman J, Grooten WJA, Rasmussen Barr E. Visual assessment of movement quality in the single leg squat test: a review and meta-analysis of inter-rater and intrarater reliability. *BMJ Open Sport & Exercise Medicine* 2019;**5**:e000541. doi:10.1136/bmjsem-2019-000541

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136bmjsem-2019-000541).

Accepted 26 May 2019

[1]Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden
[2]Allied Health Professionals Function, Functional Area Occupational Therapy and Physiotherapy, Karolinska University Hospital, Stockholm, Sweden

Correspondence to
John Ressman;
john.ressman@ki.se

ABSTRACT

Single leg squat (SLS) is a common tool used in clinical examination to set and evaluate rehabilitation goals, but also to assess lower extremity function in active people.

Objectives To conduct a review and meta-analysis on the inter-rater and intrarater reliability of the SLS, including the lateral step-down (LSD) and forward step-down (FSD) tests.

Design Review with meta-analysis.

Data sources CINAHL, Cochrane Library, Embase, Medline (OVID) and Web of Science was searched up until December 2018.

Eligibility criteria Studies were eligible for inclusion if they were methodological studies which assessed the inter-rater and/or intrarater reliability of the SLS, FSD and LSD through observation of movement quality.

Results Thirty-one studies were included. The reliability varied largely between studies (inter-rater: kappa/intraclass correlation coefficients (ICC) = 0.00–0.95; intrarater: kappa/ICC = 0.13–1.00), but most of the studies reached 'moderate' measures of agreement. The pooled results of ICC/kappa showed a 'moderate' agreement for inter-rater reliability, 0.58 (95% CI 0.50 to 0.65), and a 'substantial' agreement for intrarater reliability, 0.68 (95% CI 0.60 to 0.74). Subgroup analyses showed a higher pooled agreement for inter-rater reliability of ≤3-point rating scales while no difference was found for different numbers of segmental assessments.

Conclusion Our findings indicate that the SLS test including the FSD and LSD tests can be suitable for clinical use regardless of number of observed segments and particularly with a ≤3-point rating scale. Since most of the included studies were affected with some form of methodological bias, our findings must be interpreted with caution.

PROSPERO registration number CRD42018077822.

Summary box

What is already known?

- The single leg squat (SLS) test is an observational test for movement quality which has a widespread clinical use in assessing the lower limb.
- Visual assessment of the knee in relation to the foot is valid and reliable for use in research and clinical settings for an asymptomatic adult population.
- Due to few studies and inconsistent findings, the reliability of the SLS that assess other segments than the knee is not yet established.

What are the new findings?

- The SLS shows a moderate reliability across all types of SLS tests and is proposed as feasible and reliable in a clinical setting.
- Assessment scales with a ≤3-point rating scale shows a higher pooled agreement for inter-rater reliability compared with ≥4-point rating scales.
- The reliability is not affected by the number of observed body segments.
- Visual assessment of more than two body segments might give the clinician more information which is relevant and helpful in targeted rehabilitation.

INTRODUCTION

Visual assessment of movement is commonly used in sports medicine and aims to recognise quality of movement for identifying athletes predisposed to future injury.[1–4] For the lower extremity, a series of postural malalignments during single-limb weight bearing or landing have been characterised by excessive pelvic drop, femoral internal rotation, knee valgus, tibia internal rotation and foot pronation.[5–7] These malalignments are reportedly associated with overuse syndromes such as patellofemoral pain syndrome,[8] iliotibial pain syndrome,[9] femuro-acetabular impingement,[10] tibial stress fractures[11] and injuries such as anterior cruciate ligament injuries.[12] The single leg squat (SLS) is used to assess movement quality in the lower limb performed by squatting from a single-leg stance while the quality of the movement is observed and assessed. The SLS is described in the literature in various ways, including single-limb mini squat,[13] unilateral squat,[14] one legged squat,[15] single legged squat,[16] single leg mini squat[17] and single leg small knee bend.[18] Thus, a variety of protocols for

assessing and performing the SLS are presented,[13 14 19–22] making it difficult to define a uniform test as 'the SLS test'. Some authors propose a simple segmental approach as they assess only the relation between the foot and the knee,[13] while others propose a multisegmental approach, assessing the whole kinetic chain from the foot to the trunk.[19] In addition, assessment criteria vary,[14 22] as does performance in terms of squatting depth, arm position, support and position of the non-weight-bearing leg (ie, front, middle and back).[13 22–27] Similar to the SLS are the forward step-down (FSD) and lateral step-down (LSD) tests. These tests differ from the SLS by being performed standing on a 15–25 cm high box. Even if studies have shown kinematic and kinetic differences between various SLS[28] and in addition between SLS and FSD,[29] the movement patterns during the descendent phase are the same; flexion at the knee, hip and trunk, pelvic tilt, hip adduction and knee internal rotation and abduction.[28 29] The common denominators for these test are that they visually assess balance, stability, knee control, overall motor control, coordinated movement quality and dynamic alignment throughout the body. That is to say, the same construct with regard of lower extremity coordination patterns of the foot, knee, hip and pelvic. Based on this similarity in construct, the FSD and LSD will be included and analysed in this meta-analysis together with the SLS.

Previous literature reviews on the measurement properties of clinical tests to assess movement quality have focused on weight-bearing activities in general (eg, drop jump, tuck jump, lunge and SLS)[30 31] and showed poor to very good inter-rater and intrarater reliability. For clinical and research purposes, it is important that a test is reliable. Reliability in general is affected by factors such as the complexity of the rating scale (dichotomised or multiple-rating, number of segments assessed), the definitions of the rating criteria, the velocity of the tests and the examiner's training and clinical experience.[31 32] Besides the large between-subject variation due to biomechanical differences between individuals, an important aspect of reliability measures of these tests is the within-subject variation. Although 3D[33–37] and 2D studies[27 38–40] report joint kinematics with fair to good agreement over time, the SLS, FSD and LSD joint kinematics have not yet been adequately assessed for within-subject reliability using visual assessment.[31] To our knowledge, no review and meta-analysis have previously summarised the reliability of the SLS and included the FSD and LSD. Thus, the aim of this study was to perform a review and meta-analysis on the inter-rater and intrarater reliability of visual assessment of the SLS, including the FSD and LSD.

METHODS

The review and meta-analysis were performed according to preferred reporting items for systematic reviews and meta-analyses guidelines.[41 42]

Literature search and study selection

We conducted a systematic literature search in the CINAHL, Cochrane Library, Embase, Medline (OVID) and Web of Science databases. We used the search concepts: SLS, reproducibility of results and observer variation. The MeSH terms identified for searching Medline (OVID) were adapted in accordance with corresponding vocabularies in CINAHL and Embase. Each search concept was also complemented with relevant free-text terms and the terms were, if appropriate, truncated and/or combined with proximity operators. No language restriction was applied. Databases were searched from inception. The complete search strategies are available in online supplementary material A. The searches were performed up until 29 November 2018.

Eligibility criteria

Studies were eligible for inclusion if they were methodological studies which assessed the inter-rater and/or intrarater reliability of the SLS, FSD and LSD through observation of movement quality. No limitations were placed on participants' age, activity level or incidence of musculoskeletal disorder. Studies of inter-rater and intrarater reliability were excluded which conducted only kinematic and kinetic studies. Furthermore, studies were excluded in which the assessment was performed quantitatively through photographs where angles and degrees were calculated.

Quality assessment and risk of bias

Two authors (JR and ERB) independently assessed the studies meeting the inclusion criteria for methodological quality any disagreement was resolved by consensus discussion and with the participation of an arbitrary third researcher if required (WJAG). We used the Quality Appraisal of Reliability Studies Checklist (QAREL)[43] to assess methodological quality. QAREL is an reliable instrument specially designed to assess the quality of studies of diagnostic reliability.[44] QAREL consists of 11 items covering seven principles: sampling bias and the representativeness of subjects and raters, raters' blinding; order of raters or subject's examination; suitability of time interval among repeated measurements; application and interpretation of test and statistical analysis. Each item should be considered individually and can be answered 'yes', 'no', 'unclear' or 'not applicable'.[43]

Data extraction and synthesis

Two researchers (JR and ERB), independently and blinded to each other, screened the titles, abstracts and full papers against the inclusion and exclusion criteria. Any disagreements were resolved by consensus discussion with the third researcher if required (WJAG). The information extracted was summarised in tables, including study name, number of participants, age/gender, activity level, musculoskeletal disorders, number of examiners and their level of experience, method/test, assessment criteria and outcome/statistics. Predefined

cut-off p oints for interpretation and categorisation of results were used. For the kappa coefficient, first order of agreement coefficient (AC1) and intraclass correlation coefficients (ICC), the Landis and Koch[45] classification for agreement was used; κ/ICC/AC1:<0.00 = poor; κ/ICC/AC1: 0.00–0.20 = slight; κ/ICC/AC1: 0.21–0.40 = fair; κ/ICC/AC1: 0.41–0.60 = moderate; κ/ICC/AC1: 0.61–0.80 = substantial and κ/ICC/AC1: 0.81–1.0 = almost perfect.

We pooled data and conducted two separate meta-analyses for inter-rater and intrarater reliability across all studies. Reliability estimates (ICC, kappa and AC1) and sample size values were extracted from each study and transformed to Fisher's z scale.[46–49] Transformation to Fisher's z is used in correlational meta-analyses to account for the non-normal distribution in these types of statistics.[46–49] A random-effect model was used due to expected heterogeneity between studies, the between-studies and total between-subgroup effect size heterogeneity was conducted following the transformation to Fisher's z using the Q test and the result was expressed as I^2 statistics. To aid in the interpretation of the results, Fisher's z values were then converted back to reliability estimate values after completing the meta-analyses. The effect size was expressed as the pooled agreement of ICC, kappa and AC1 with 95% CI and for all outcome measures, the critical value to reject H_0 was set to 0.05. All statistical analyses were completed using comprehensive meta-analysis V.3.[50]

For the meta-analyses, three choices were made. First, when more than one reliability data were presented for the same rating, a mean value was calculated for multiple examiners (where the experienced examiners were chosen), dominant/non-dominant leg, rating of different segments (ie, hip and knee) and for school children in third and seventh grades.[30] Second, in two of the included studies,[14 23] different assessment methods for the same test were presented and in these cases, the method most conform with the other included methods was chosen. Third, to include reliability data mostly with the same measurement units; plain kappa was chosen before weighted kappa, prevalence-adjusted bias-adjusted kappa, generalised kappa and weighted generalised kappa.

We conducted subgroup analyses to study differences in reliability due to different approaches in the assessment criteria; (1) on the number of segments rated uniseg-mental/bisegmental approach containing one or two segments versus a multisegmental approach containing ≥3 segments (2) the rater's rating scale (≤3-point vs ≥4-point rating scales).

As a final step, we conducted four sensitivity analyses to test the robustness of our results.

1. To investigate the importance of study quality, we conducted an analysis in which studies assessed with 'no' according to QAREL were removed.
2. To investigate if exclusion of assessment methods not considered conform to other included assessment methods changes our results, those methods were included in the meta-analysis.[14 23]
3. To investigate if exclusion of all FSD and LSD tests changed our pooled results, we performed meta-analyses excluding these tests. The tests described in Crossley *et al*[19] and McKeown *et al*[24] were considered as an FSD, thus being described as such, even if they were presented as SLS by the authors.
4. To confirm that our findings were not driven by any single study, a leave-one-out sensitivity analysis was also performed by removing one study at a time, iteratively.

RESULTS

Study selection

The literature search elicited 5230 references of which 2367 were duplicates and another 2800 were excluded after screening titles and abstracts (figure 1). In total, 68 studies were reviewed in full text after further citation tracking of the references lists of included studies. We included 31 studies in the review, while 37 studies were excluded for one of the following reasons: kinematic/kinetic studies (n=9), quantitative measures of SLS (n=5), no methodological studies (n=4), not evaluating a test similar to SLS (n=17) or composite results of more than one test (n=2). Fifteen of the included studies investigated inter-rater and intrarater reliability,[14 16 18 19 22–24 51–58] 14 investigated inter-rater reliability[13 15 17 20 21 25 26 59–65] and two investigated intrarater reliability.[27 66]

Risk of bias within studies

The methodological quality of the included studies, assessed with QAREL, is presented in table 1. Seven studies were assessed as not fulfilling item 11[19 20 22 24 53 59 66] which evaluates the statistical analysis of reliability and one study[56] did not fulfil item 8 concerning the order in which raters' or subjects are examined. All studies were assessed with one or more 'uncertain' concerning examiner blinding.

Study characteristics

The specific study characteristics are presented in online supplementary materials B and C.

Subjects

Altogether, the 31 studies included 1136 subjects (454 female, 360 males and 322 of unknown gender) with an age range from 9 to 89 years; 65% of the subjects were healthy and active or were athletes between 18 and 37 years old. Five studies investigated symptomatic subjects with hip osteoarthrosis,[51] patellofemoral pain syndrome,[26] anterior cruciate ligament injury,[62] knee osteoarthritis[58 63] and three studies investigated healthy children aged from 9 to 16 years.[17 18 55]

Examiners

The examiners in the studies comprised 272 certified physiotherapists with different clinical experiences; 45 physiotherapy students or non-clinician physiotherapists;

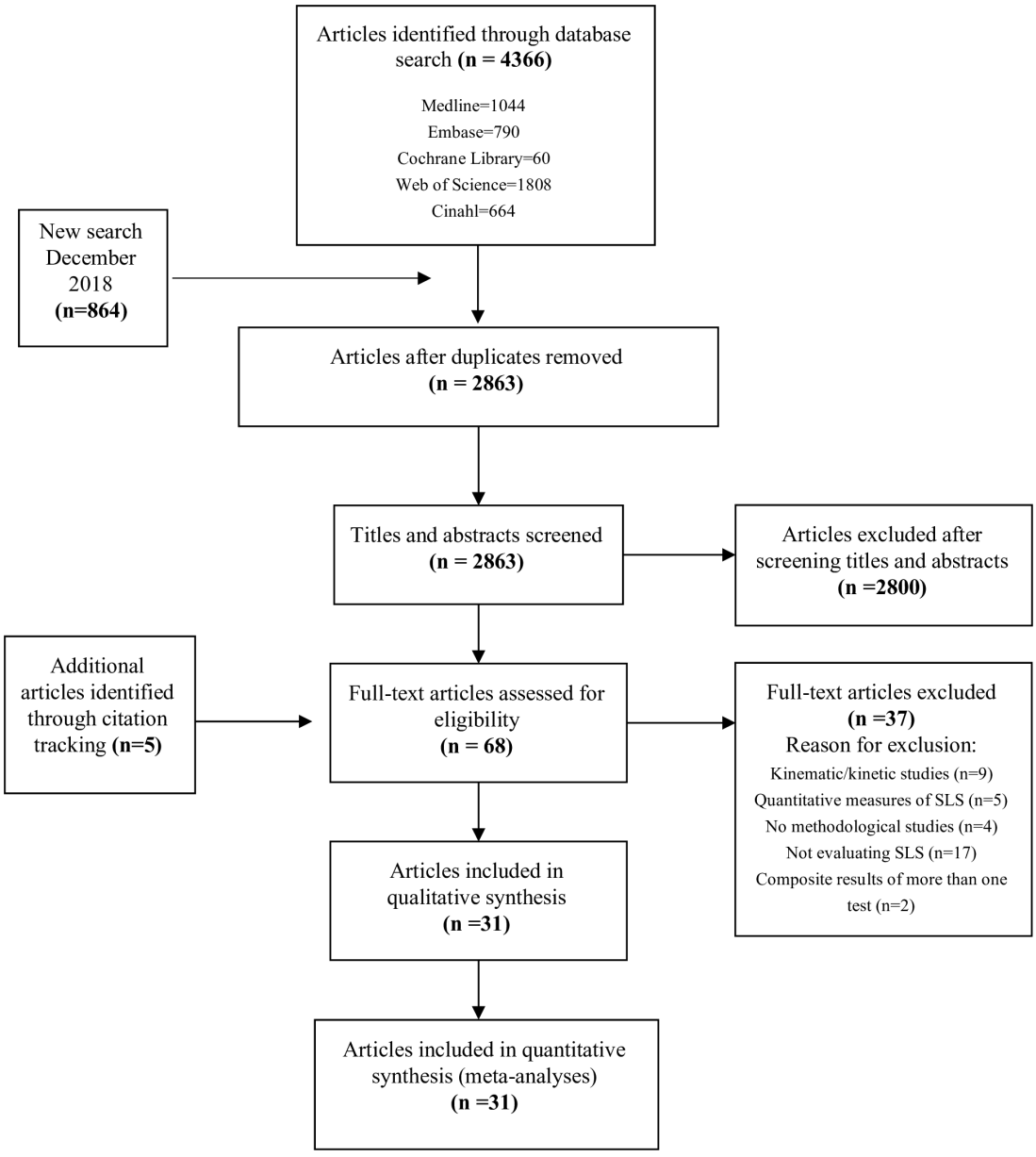

Figure 1 Flow chart of inclusion process.

eight athletic trainers; six strength and conditioning coaches; eight physicians; four orthopaedic surgeons and eight examiners of unknown profession.

Tests

The 31 studies covered 34 tests and presented a variety of different tests or the same tests with variations in name, protocols and methods of performance (online supplementary material B). Three studies[14 54 56] presented two separate tests. Twelve studies investigated the SLS.[16 19 20 22 23 27 52 53 56 57 59 61] None of these presented the SLS with identical protocols except for Stensrud *et al*[27] and Raisanen *et al*,[53] who described the test in a similar way. Six tests were named as the LSD[14 26 54 60 64 65] and were similar in performance but used boxes of different heights, while two tests were named FSD[21 25] and differed in arm position. Furthermore, three tests were named single-limb mini squat,[13 62 63] two tests were named unilateral squat[14 54] and eight tests used different names: single leg mini squat,[17] SLS off a box,[24] single leg small knee bend,[18] small knee bend,[56] one-leg squat test,[66] small squat on one-leg stance,[58] small SLS[51] and one-legged squat.[15]

Assessment criteria

In seven studies, the visual assessment was scored by a 2-point rating scale (dichotomous),[13 18 23 55 56 59 66] five studies used a mixed 2-point and 3-point rating scale,[25 26 60 64 65] nine studies used a 3-point rating scale[16 19 21 24 27 51 53 57 58] and another nine used a 4-point rating scale,[14 15 17 20 52 54 61–63] one study used a 10-point rating scale.[22] Most of the studies used a multisegmental approach (≥3 observed segments); 21 studies observed four segments or more,[14 15 17 20–26 51 54 56 58 60–66] four studies observed three

Table 1 Methodological quality and risk of bias of included studies assessed with the Quality Appraisal of Reliability Studies*

Study	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Ageberg *et al*[13]	**Y**	**Y**	**Y**	NA	U	U	U	NA	**Y**	**Y**	**Y**
Barker-Davis *et al*[56]	**Y**	**Y**	**Y**	U	U	NA	U	N	**Y**	**Y**	**Y**
Chmielewski *et al*[14]	**Y**	**Y**	U	U	NA	U	U	**Y**	**Y**	**Y**	**Y**
Cornell *et al*[66]	**Y**	**Y**	NA	U	NA	U	U	U	Y	Y	N
Crossley *et al*[19]	**Y**	**Y**	U	U	**Y**	U	U	**Y**	**Y**	**Y**	N
Di Mattia *et al*[20]	**Y**	**Y**	U	NA	U	U	U	NA	**Y**	**Y**	N
Edmondston *et al*[59]	**Y**	**Y**	U	NA	U	U	U	**Y**	**Y**	**Y**	N
Friedrich *et al*[61]	**Y**	**Y**	**Y**	NA	NA	NA	**Y**	U	U	**Y**	**Y**
Frohm *et al*[15]	**Y**	**Y**	**Y**	NA	NA	**Y**	U	NA	**Y**	**Y**	**Y**
Gianola *et al*[57]	**Y**	**Y**	**Y**	U	U	NA	U	U	**Y**	**Y**	**Y**
Harris-Hayes *et al*[16]	**Y**	**Y**	**Y**	U	**Y**	U	U	**Y**	**Y**	**Y**	**Y**
Herman *et al*[21]	**Y**	**Y**	U	NA	**Y**	U	U	NA	**Y**	**Y**	**Y**
Junge *et al*[17]	**Y**	**Y**	U	NA	NA	U	U	NA	**Y**	**Y**	**Y**
Kaukinen *et al*[58]	**Y**	**Y**	**Y**	**Y**	NA	**Y**	U	**Y**	**Y**	**Y**	**Y**
Kennedy *et al*[23]	**Y**	**Y**	U	U	NA	U	U	U	**Y**	**Y**	**Y**
Lenzlinger-Asprion *et al*[51]	**Y**	**Y**	**Y**	U	NA	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
McKeown *et al*[24]	**Y**	**Y**	U	U	NA	U	U	NA	**Y**	**Y**	N
Nae *et al*[30]	**Y**	**Y**	**Y**	NA	NA	U	U	U	**Y**	**Y**	**Y**
Park *et al*[25]	**Y**	**Y**	**Y**	NA	NA	U	U	**Y**	**Y**	**Y**	**Y**
Piva *et al*[26]	**Y**	**Y**	**Y**	NA	NA	U	U	**Y**	U	**Y**	**Y**
Poulsen *et al*[52]	**Y**	**Y**	**Y**	**Y**	U	U	U	U	**Y**	**Y**	**Y**
Rabin *et al*[64]	**Y**	**Y**	**Y**	NA	NA	U	U	NA	**Y**	**Y**	**Y**
Rabin *et al*[65]	**Y**	**Y**	**Y**	NA	NA	U	U	NA	**Y**	**Y**	**Y**
Räisänen *et al*[53]	**Y**	**Y**	U	**Y**	U	U	U	**Y**	**Y**	**Y**	N
Stensrud *et al*[27]	**Y**	**Y**	NA	U	U	U	U	U	U	**Y**	**Y**
Teyhen *et al*[60]	**Y**	**Y**	**Y**	NA	NA	U	U	**Y**	**Y**	**Y**	**Y**
Van Mastrigt *et al*[63]	**Y**	**Y**	**Y**	NA	U	**Y**	U	**Y**	**Y**	U	**Y**
Weeks *et al*[22]	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	U	**Y**	**Y**	**Y**	N
Weir *et al*[54]	**Y**	**Y**	U	U	NA	U	U	**Y**	**Y**	**Y**	**Y**
Whatman *et al*[18]	**Y**	**Y**	U	U	**Y**	U	U	**Y**	**Y**	**Y**	**Y**
Örtqvist *et al*[55]	**Y**	**Y**	**Y**	U	NA	U	U	**Y**	U	**Y**	**Y**

*Assesses study quality based on 11 items. Items: 1. Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied? 2. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied? 3. Were raters blinded to the findings of other raters during the study? 4. Were raters blinded to their own prior findings of the test under evaluation? 5. Were raters blinded to the results of the accepted reference standard or the disease status for the target disorder (or variable) being evaluated? 6. Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design? 7. Were raters blinded to additional cues that were not part of the test? 8. Was the order of examination varied? 9. Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time-interval among repeated measures? 10. Was the test applied correctly and interpreted appropriately? 11. Were appropriate statistical measures of agreement used?
N, no; NA, not applicable; U, unclear; Y, yes (marked in bold).

segments,[18 19 52 57] five observed two segments[13 16 27 53 55] and one study observed only one segment.[59]

Synthesis of results

The ICC, AC1 and kappa values of the included studies are shown in online supplementary material C.

Inter-rater reliability

In total, 29 studies reported on inter-rater reliability and varied between 'slight' and 'almost perfect' (κ=0.00–0.95/ICC=0.39–0.71) (online supplementary material C). Twenty-two of these presented inter-rater agreement varying between 'moderate' and 'almost perfect' (κ and

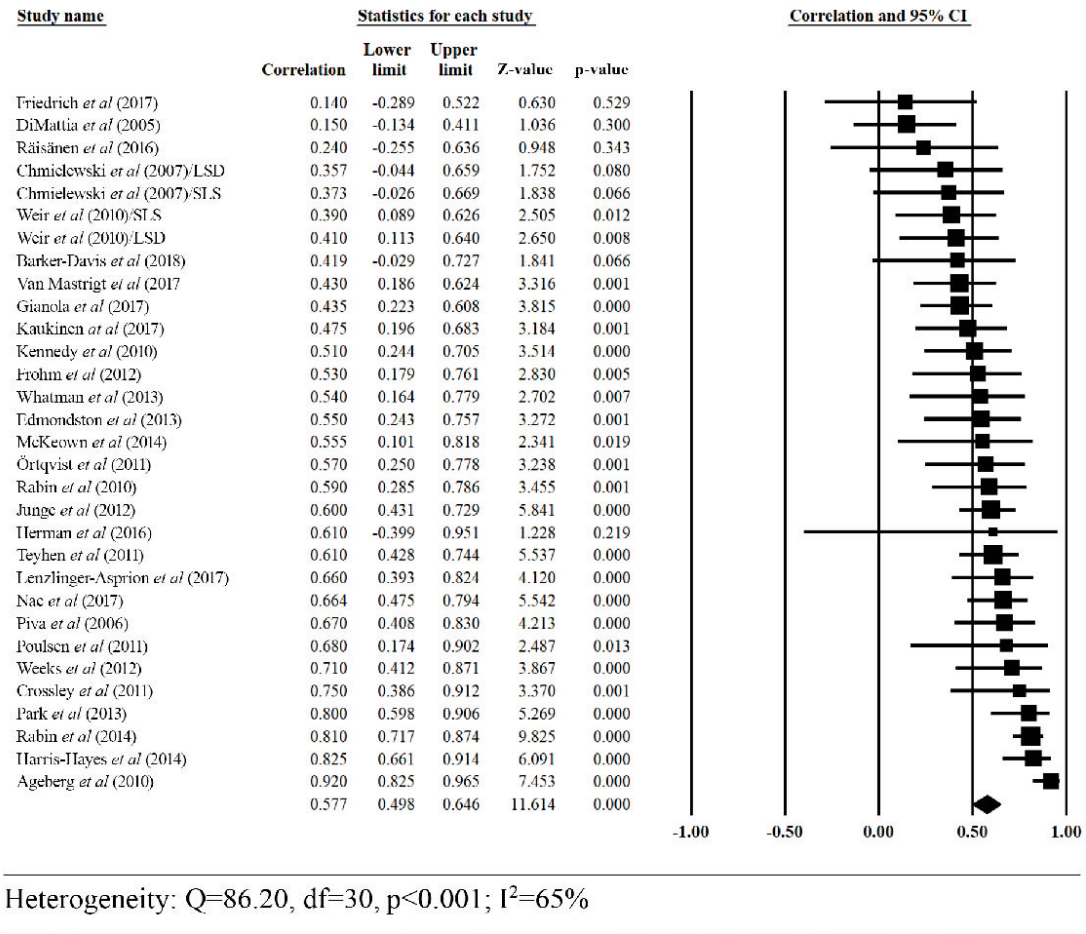

Figure 2 Forest plot and the pooled agreement coefficient of studies on the agreement coefficient (ICC, kappa and AC1) for inter-rater reliability of the single-leg squat in a random effect model.

ICC≥ 0.41).[13 15–19 21 22 24–26 51 52 54–56 58–60 62 64 65] The pooled agreement for ICC, kappa and AC1 was 0.58 (95% CI 0.50 to 0.65), indicating a 'moderate' agreement (figure 2). The test for heterogeneity was significant (Q=86.20, df=30, p<0.001) and the I^2 statistics reported that 65% of the variability was attributed to heterogeneity.

Intrarater reliability

Seventeen of the included studies investigated intra-rater reliability and varied between 'slight' and 'almost perfect' (κ=0.13–1.00/ICC=0.49–0.81) (online supplementary material C). Twelve studies presented intrarater agreement varying between 'moderate' and 'almost perfect' (κ and ICC ≥ 0.41).[16 18 19 22 24 51 54–58 66] The pooled agreement was 0.68 (95% CI 0.60 to 0.74), indicating a 'substantial' agreement (figure 3). The test for heterogeneity was significant (Q=38.46, df=18, p=0.003) and the I^2 statistics reported that 53% of the variability was attributed to heterogeneity.

Subgroup analysis

Subgroup analyses relating to the assessment criteria; segmental approach and rating scale are presented in online supplementary material D-G. Subgroup analysis showed no significant difference for inter-rater reliability between unisegmental/bisegmental approach and multisegmental approach 0.62 (95% CI 0.44 to 0.76) versus 0.57 (95% CI 0.47 to 0.65), p=0.56. The pooled agreement for intrarater reliability was 0.72 (95% CI 0.56 to 0.82) for the unisegmental/bisegmental approach and 0.66 (95% CI 0.58 to 0.74) for the multisegmental approach (p=0.53). For rating scales, the subgroup analysis showed a significant difference with a pooled agreement for inter-rater reliability of 0.64 (95% CI 0.56 to 0.71) for the ≤3-point rating scale versus 0.47 (95% CI 0.33 to 0.58) for the ≥4-point rating scale (p=0.016). For intrarater reliability, the pooled agreement was 0.71 (95% CI 0.62 to 0.77) for the ≤3-point rating scale and 0.60 (95% CI 0.44 to 0.73) for the ≥4-point rating scale (p=0.18).

Sensitivity analyses

Seven studies[19 20 22 24 53 59 66] did not fulfil QAREL item 11 and one study[56] did not fulfil item 8. Sensitivity analysis on the importance of study quality showed that the pooled agreement for inter-rater reliability slightly increased to 0.60 (95% CI 0.51 to 0.67), while the intrarater reliability decreased to 0.62 (95% CI 0.53 to 0.71) when those eight studies were eliminated from the meta-analyses. Three assessment methods in two studies[14 23] were initially excluded from the meta-analyses in order to achieve

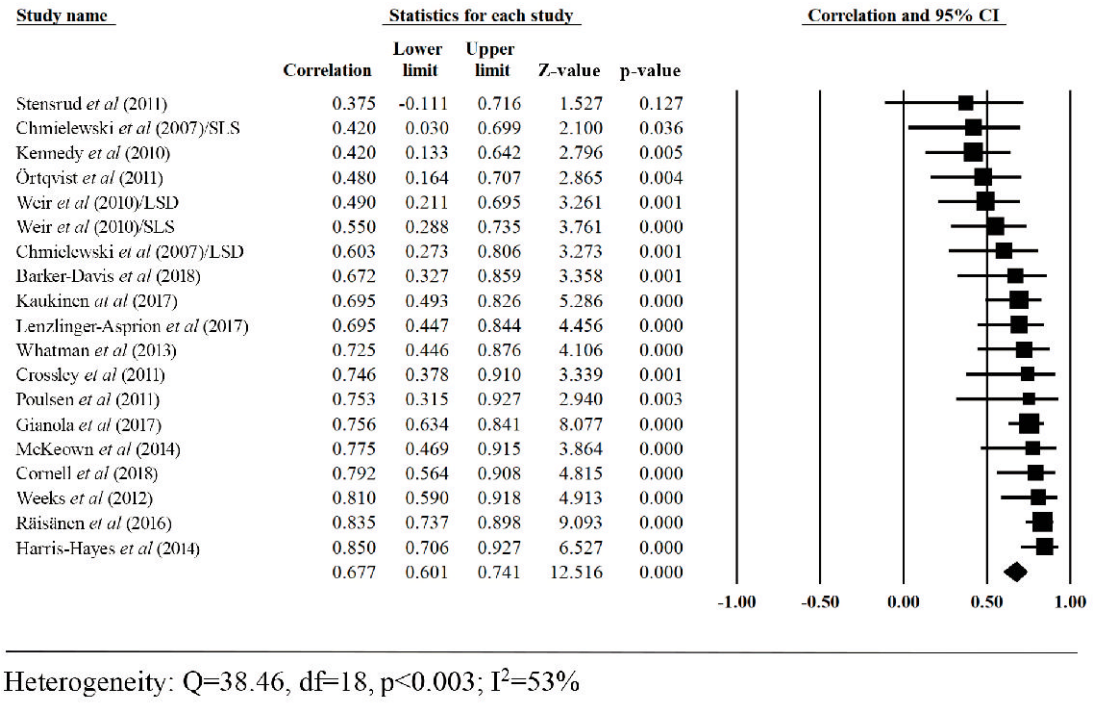

Figure 3 Forest plot and the pooled agreement coefficient of studies on the agreement coefficient (ICC, kappa nd AC1) for intrarater reliability of the single-leg squat in a random effect model.

conformity. Sensitivity analyses on including these three assessment methods showed a slightly overall decreased pooled agreement of 0.56 (95% CI 0.48 to 0.63) for inter-rater reliability and 0.65 (95% CI 0.57 to 0.72) for intrarater reliability.

Six of the included studies[14 26 54 60 64 65] presented LSD tests and four studies[19 21 24 25] FSD tests. The sensitivity analyses showed that the pooled agreement for inter-rater reliability slightly decreased to 0.55 (95% CI 0.45 to 0.63) and for intrarater reliability slightly increased to 0.69 (95% CI 0.60 to 0.76) when all FSD and LSD tests were excluded from the meta-analyses. The same small changes of the pooled results were seen when only the LSD tests were excluded; inter-rater reliability of 0.57 (95% CI 0.48 to 0.65) and intrarater reliability of 0.69 (95% CI 0.61 to 0.76). When the FSD tests were excluded, an inter-rater reliability of 0.56 (95% CI 0.48 to 0.64) and an intrarater reliability of 0.67 (95% CI 0.59 to 0.74) were seen. The leave-one-out sensitivity analysis indicated that the pooled agreement remained 'moderate' for inter-rater reliability and 'substantial' for intrarater reliability despite removing any single study from the analysis.

DISCUSSION

We conducted a review and meta-analyses of the inter-rater and intrarater reliability for the visual assessment of the SLS, including the LSD and FSD. For both the inter-rater and intrarater reliability, most studies found a 'moderate' to 'almost perfect' agreement. The meta-analyses showed a pooled agreement for inter-rater reliability of 0.58 (95% CI 0.50 to 0.65), indicating a 'moderate' agreement while the intrarater reliability was somewhat higher 0.68 (95% CI 0.60 to 0.74), indicating a 'substantial' agreement. Sensitivity analyses did not change the pooled results. Subgroup analyses showed no differences regarding unisegmental/bisegmental versus a multisegmental approach for both inter-rater and intrarater reliability, while the inter-rater reliability of a≤3-point rating scale was significantly greater than the ≥4-point rating scale. There were, however, no difference detected concerning intrarater reliability.

Previous literature reviews have focused on weight-bearing activities in general (ie, drop jump, tuck jump, lunge and SLS),[30 31] the validity/kinematics of such tests[67] or modifiable factors associated with knee abduction during weight-bearing activities.[68] Nae *et al*[30] concluded that visual assessment of the knee in relation to the foot is valid and reliable for use in research and clinical settings for an asymptomatic adult population. In concordance with this, Whatman *et al*[31] showed acceptable reliability for various SLS across a range of ages, using a dichotomous rating of the knee in relation to the foot. Further, Nae *et al*[30] and Whatman *et al*[31] concluded that clearly described assessment criteria, a dichotomous rating scale and a visual assessment on video increased the reliability. This is echoed by our findings, which in addition to previous reviews[30 31] included 15 additional studies.[21–24 51 53 56–58 61–66] Yet, none of the additional studies focused solely on the relation between the knee and foot, as most of them used a multisegmental approach.[21–24 51 56–58 61–66] Nae *et al*[30] stated in their review that the reliability of tests that assess other segments than the knee is not yet established, due to few studies and inconsistent findings.[30] The present review, however, shows that SLS tests using

either an unisegmental/bisegmental or multisegmental approach exhibit an acceptable inter-rater and intrarater reliability and that most of the included studies which used a multisegmental approach exhibited an inter-rater reliability ranging from 'moderate' to 'almost perfect' (κ/ICC >0.41). This was also supported by the subgroup analysis which showed no differences between the unisegmental/bisegmental and multisegmental ratings. Moreover, Whatman *et al*[31] found that assessment using more complex ratings, such as 3-point and 4-point rating scales or rating multiple segments, has acceptable reliability in some studies but are generally not considered reliable enough. Whatman *et al*[31] in addition proposed that more complex methods warrant further investigation as they may provide clinicians with information that could be relevant to clinical decision making. However, our subgroup analyses show that 2-point or 3-point rating scales versus ≥4-point rating scale seem to be superior. Hence, our results show that observer rating regardless of number of assessed segments, and furthermore ratings on a ≥3-point rating scale show an acceptable agreement. This indicate that such tests may be of clinical use.

Different cut-off scores for ICC and kappa values exist in the literature; for example, Streiner *et al*[69] recommend a kappa value of 0.60–0.75 for tests to be considered reliable. Our findings from the meta-analyses showed that the intrarater reliability across all studies and those studies with ≤3-point rating scales exceeded 0.60, but the pooled agreement coefficient for the inter-rater reliability was just below 0.60. On the other hand, previous studies on reliability suggest that a lower cut-off score (κ>0.40) might be considered sufficient for a test to be used in clinical work.[70–73] We consider this reasonable, as examiners will have different experiences and act in different settings and those being assessed will vary. Hence, we believe that we can conclude that these tests are reliable enough to be of use in clinical practice.

The methodological quality of the included studies may be questioned, as all studies were assessed as 'uncertain' for one or more items, indicating an information gap due to insufficient information provided in the study. In most cases, items assessed as 'uncertain' were related to examiner blinding. When assessing the risk of bias, it cannot be assumed that the examiners were blinded if this is not clearly stated. Future research studies should therefore ensure that examiners are blinded and clearly state this in the methodological section. In addition, seven[19 20 22 24 53 59 66] studies did not fulfil QAREL item 11 and one study did not fulfil item 8.[56] Regardless, the sensitivity analysis of methodological study quality showed that the pooled agreement stayed above 'moderate' when those seven studies were where eliminated from the meta-analyses.

Our meta-analyses found a moderate heterogeneity[74] between included studies (I^2=53%–65%) suggesting a great variability across all included studies which also has been reported in previous reviews.[30 31] Included studies varied in performance and assessment protocols, study populations and examiners' experiences, suggesting need for further standardisation of testing.

A strength of the present study is its extensive literature search and robustness of the employed methodology. Another strength is the performance of pooled analyses, including subgroup analyses, summarising more than 30 studies on SLS tests similar in performance. To merge various tests in one review may be considered advantageous as it presents the opportunity to compare multiple results from different studies. On the other hand, one could argue that the SLS, FSD and LSD differ and therefore cannot be compared due to the variation in their biomechanical effects in kinematic and kinetic demands.[28 29] Nevertheless, a sensitivity analysis excluding all FSD and LSD tests in our meta-analysis showed only a slight change in the pooled agreement which confirms the robustness of our results and indicates that the visual assessment of a SLS, regardless of stepping-down from a box or performing a SLS standing on the floor shows moderate to substantial reliability. A limitation of the present review is its decreased generalisability to populations other than healthy/active people aged 18 to 37 years, even though the present review includes five studies involving symptomatic subjects,[26 51 58 62 63] three studies involving healthy children aged nine to 16 years[17 18 55] and three studies involving older people aged between 55 and 89 years.[51 58 63] Further, different correlation statistics were merged, thus many of the studies included used different kappa statistics and ICC models, or did not report the ICC model used, which could have had implications for the pooled agreement estimates. For the meta-analyses, some choices were needed to be made if more than one reliability measure was presented for the same rating, when different assessment methods were presented in the same study and concerning the choice of the kappa statistics. However, we considered this necessary for the data processing and this methodology has previously been reported.[30] Finally, there is always a risk that a study has been missed out due to poor indexing of studies.

CONCLUSION

Our results indicate that the SLS test including the FSD and LSD tests are feasible and reliable, regardless of whether a unisegmental/bisegmental or a multisegmental approach is used. Our findings show a 'moderate' reliability in assessment of the SLS, indicating that the test is suitable for use in clinical work regardless of number of observed segments and particularly with a ≤3-point rating scale. Since most of the included studies are affected with some methodological bias, our findings must be interpreted with caution. Future studies using more robust methodological standardisation of the test performance are warranted.

Acknowledgements The authors would like to thank Magdalena Svanberg and Gun Brit Knutssön, librarians at Karolinska Institutet University Library, for their help with the development of the search strategies and database search.

Contributors JR contributed to the design of the study, and was responsible for collecting, analysing and interpreting the data and for drafting the manuscript together with ERB. ERB contributed to the conception and design of the study, undertook analysis and interpretation of data, drafted the manuscript together with JR and provided feedback on drafts of the manuscript. WJAG contributed in analysis and interpretation of data and provided feedback on drafts of the manuscript. All three authors read and approved the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests Competing interest.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movements as an assessment of function - part 1. *N Am J Sports Phys Ther* 2006;1:62–72.
2. Cook G, Burton L, Hoogenboom B. Pre-participation screening: the use of fundamental movements as an assessment of function - part 2. *N Am J Sports Phys Ther* 2006;1:132–9.
3. McCall A, Carling C, Nedelec M, *et al*. Risk factors, testing and preventative strategies for non-contact injuries in professional football: current perceptions and practices of 44 teams from various premier leagues. *Br J Sports Med* 2014;48:1352–7.
4. McCunn R, Aus der Fünten K, Fullagar HHK, *et al*. Reliability and association with injury of movement screens: a critical review. *Sports Med* 2016;46:763–81.
5. Hewett TE, Myer GD, Ford KR, *et al*. Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *Am J Sports Med* 2005;33:492–501.
6. Hollman JH, Ginos BE, Kozuchowski J, *et al*. Relationships between knee valgus, hip-muscle strength, and hip-muscle recruitment during a single-limb step-down. *J Sport Rehabil* 2009;18:104–17.
7. Powers CM. The influence of altered lower-extremity kinematics on patellofemoral joint dysfunction: a theoretical perspective. *J Orthop Sports Phys Ther* 2003;33:639–46.
8. Herrington L. Knee valgus angle during single leg squat and landing in patellofemoral pain patients and controls. *The Knee* 2014;21:514–7.
9. Aderem J, Louw QA. Biomechanical risk factors associated with iliotibial band syndrome in runners: a systematic review. *BMC Musculoskelet Disord* 2015;16.
10. Botha N, Warner M, Gimpel M, *et al*. Movement patterns during a small knee bend test in Academy footballers with femoroacetabular impingement (FAI). *Health Sciences Working Papers* 2014;1:1–24.
11. Milner CE, Hamill J, Davis IS. Distinct hip and rearfoot kinematics in female runners with a history of tibial stress fracture. *J Orthop Sports Phys Ther* 2010;40:59–66.
12. Hewett TE, Myer GD, Ford KR, *et al*. Mechanisms, prediction, and prevention of ACL injuries: cut risk with three sharpened and validated tools. *J Orthop Res* 2016;34:1843–55.
13. Ageberg E, Bennell KL, Hunt MA, *et al*. Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. *BMC Musculoskelet Disord* 2010;11.
14. Chmielewski TL, Hodges MJ, Horodyski M, *et al*. Investigation of clinician agreement in evaluating movement quality during unilateral lower extremity functional tasks: a comparison of 2 rating methods. *J Orthop Sports Phys Ther* 2007;37:122–9.
15. Frohm A, Heijne A, Kowalski J, *et al*. A nine-test screening battery for athletes: a reliability study. *Scand J Med Sci Sports* 2012;22:306–15.
16. Harris-Hayes M, Steger-May K, Koh C, *et al*. Classification of lower extremity movement patterns based on visual assessment: reliability and correlation with 2-Dimensional video analysis. *Journal of Athletic Training* 2014;49:304–10.
17. Junge T, Balsnes S, Runge L, *et al*. Single leg mini squat: an inter-tester reproducibility study of children in the age of 9-10 and 12-14 years presented by various methods of kappa calculation. *BMC Musculoskelet Disord* 2012;13.
18. Whatman C, Hume P, Hing W. The reliability and validity of physiotherapist visual rating of dynamic pelvis and knee alignment in young athletes. *Physical Therapy in Sport* 2013;14:168–74.
19. Crossley KM, Zhang W-J, Schache AG, *et al*. Performance on the single-leg squat task indicates hip abductor muscle function. *Am J Sports Med* 2011;39:866–73.
20. DiMattia MA, Livengood AL, Uhl TL, *et al*. What are the validity of the Single-Leg-Squat test and its relationship to Hip-Abduction strength? *J Sport Rehabil* 2005;14:108–23.
21. Herman G, Nakdimon O, Levinger P, *et al*. Agreement of an evaluation of the Forward-Step-Down test by a broad cohort of clinicians with that of an expert panel. *J Sport Rehabil* 2016;25:227–32.
22. Weeks BK, Carty CP, Horan SA. Kinematic predictors of single-leg squat performance: a comparison of experienced physiotherapists and student physiotherapists. *BMC Musculoskelet Disord* 2012;13.
23. Kennedy MD, Burrows L, Parent E. Intrarater and interrater reliability of the single-leg squat test. *Athletic Therapy Today* 2010;15:32–6.
24. McKeown I, Taylor-McKeown K, Woods C, *et al*. Athletic ability assessment: a movement assessment protocol for athletes. *Int J Sports Phys Ther* 2014;9.
25. Park K-M, Cynn H-S, Choung S-D. Musculoskeletal predictors of movement quality for the forward step-down test in asymptomatic women. *J Orthop Sports Phys Ther* 2013;43:504–10.
26. Piva SR, Fitzgerald K, Irrgang JJ, *et al*. Reliability of measures of impairments associated with patellofemoral pain syndrome. *BMC Musculoskelet Disord* 2006;7.
27. Stensrud S, Myklebust G, Kristianslund E, *et al*. Correlation between two-dimensional video analysis and subjective assessment in evaluating knee control among elite female team handball players. *Br J Sports Med* 2011;45:589–95.
28. Khuu A, Foch E, Lewis CL. Not all single leg SQUATS are equal: a biomechanical comparison of three variations. *Int J Sports Phys Ther* 2016;11:201–11.
29. Lewis CL, Foch E, Luko MM, *et al*. Differences in lower extremity and trunk kinematics between single leg squat and step down tasks. *Plos One* 2015;10:e0126258.
30. Nae J, Creaby MW, Cronström A, *et al*. Measurement properties of visual rating of postural orientation errors of the lower extremity - A systematic review and meta-analysis. *Phys Ther Sport* 2017;27:52–64.
31. Whatman C, Hume P, Hing W. The reliability and validity of visual rating of dynamic alignment during lower extremity functional screening tests: a review of the literature. *Phys Ther Rev* 2015;20:210–24.
32. Knudson D. What can professionals qualitatively analyze? *J Phys Educ Recreat Dance* 2000;71:19–23.
33. Alenezi F, Herrington L, Jones P, *et al*. The reliability of biomechanical variables collected during single leg squat and landing tasks. *J Electromyogr Kinesiol* 2014;24:718–21.
34. Earl JE, Hertel J, Denegar CR. Patterns of dynamic malalignment, muscle activation, joint motion, and patellofemoral-pain syndrome. *J Sport Rehabil* 2005;14:216–33.
35. Earl JE, Monteiro SK, Snyder KR. Differences in lower extremity kinematics between a bilateral drop-vertical jump and a single-leg step-down. *J Orthop Sports Phys Ther* 2007;37:245–52.
36. Whatman C, Hing W, Hume P. Kinematics during lower extremity functional screening tests--are they reliable and related to jogging? *Phys Ther Sport* 2011;12:22–9.
37. Whatman C, Hume P, Hing W. Kinematics during lower extremity functional screening tests in young athletes - are they reliable and valid? *Phys Ther Sport* 2013;14:87–93.
38. Levinger P, Gilleard W, Coleman C. Femoral medial deviation angle during a one-leg squat test in individuals with patellofemoral pain syndrome. *Phys Ther Sport* 2007;8:163–8.
39. Munro A, Herrington L, Carolan M. Reliability of 2-Dimensional video assessment of frontal-plane dynamic knee valgus during common athletic screening tasks. *J Sport Rehabil* 2012;21:7–11.
40. Willson JD, Ireland ML, Davis I. Core strength and lower extremity alignment during single leg squats. *Med Sci Sports Exerc* 2006;38:945–52.
41. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
42. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
43. Lucas NP, Macaskill P, Irwig L, *et al*. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol* 2010;63:854–61.

44. Lucas N, Macaskill P, Irwig L, *et al*. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol* 2013;13.
45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
46. Borenstein M, Hedges LV, Higgins JPT, *et al*. *Introduction to Meta-Analysis*. Wiley, 2011.
47. Botella J, Suero M, Gambara H. Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychol Methods* 2010;15:386–97.
48. Cuchna JW, Hoch MC, Hoch JM. The interrater and intrarater reliability of the functional movement screen: a systematic review with meta-analysis. *Phys Ther Sport* 2016;19:57–65.
49. Ottenbacher KJ, Hsu Y, Granger CV, *et al*. The reliability of the functional independence Measure: a quantitative review. *Arch Phys Med Rehabil* 1996;77:1226–32.
50. Pierce CA. Software Review: Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R 2006. Comprehensive Meta-Analysis (Version 2.2.027) [Computer software]. Englewood, NJ: Biostat. In. Vol 11. Los Angeles, CA2008:188-191.
51. Lenzlinger-Asprion R, Keller N, Meichtry A, *et al*. Intertester and intratester reliability of movement control tests on the hip for patients with hip osteoarthritis. *BMC Musculoskelet Disord* 2017;18.
52. Poulsen DR, James CR. Concurrent validity and reliability of clinical evaluation of the single leg squat. *Physiother Theory Pract* 2011;27:586–94.
53. Räisänen A, Pasanen K, Krosshaug T, *et al*. Single-Leg squat as a tool to evaluate young athletes' frontal plane knee control. *Clin J Sport Med* 2016;26:478–82.
54. Weir A, Darby J, Inklaar H, *et al*. Core stability: inter- and intraobserver reliability of 6 clinical tests. *Clin J Sport Med* 2010;20:34–8.
55. Örtqvist M, Moström EB, Roos EM, *et al*. Reliability and reference values of two clinical measurements of dynamic and static knee position in healthy children. *Knee Surg Sports Traumatol Arthrosc* 2011;19:2060–6.
56. Barker-Davies RM, Roberts A, Bennett AN, *et al*. Single leg squat ratings by clinicians are reliable and predict excessive hip internal rotation moment. *Gait Posture* 2018;61:453–8.
57. Gianola S, Castellini G, Stucovitz E, *et al*. Single leg squat performance in physically and non-physically active individuals: a cross-sectional study. *BMC Musculoskelet Disord* 2017;18.
58. Kaukinen PT, Arokoski JP, Huber EO, *et al*. Intertester and intratester reliability of a movement control test battery for patients with knee osteoarthritis and controls. *J Musculoskelet Neuronal Interact* 2017;17:197–208.
59. Edmondston S, Leo Y, Trant B, *et al*. Symmetry of trunk and femoro-pelvic movement responses to single leg loading tests in asymptomatic females. *Man Ther* 2013;18:231–6.
60. Teyhen DS, Shaffer SW, Lorenson CL, *et al*. Reliability of lower quarter physical performance measures in healthy service members. *US Army Med Dep J* 2011:37–49.
61. Friedrich J, Brakke R, Akuthota V, *et al*. Reliability and practicality of the core score: four dynamic core stability tests performed in a physician office setting. *Clin J Sport Med* 2017;27:409–14.
62. Nae J, Creaby MW, Nilsson G, *et al*. Measurement properties of a test battery to assess postural orientation during functional tasks in patients undergoing anterior cruciate ligament injury rehabilitation. *J Orthop Sports Phys Ther* 2017;47:863–73.
63. Mastrigt Nvan, Naili JE, Broström EW, *et al*. Inter-rater reliability of movement quality during single limb mini-squat test in adults with knee osteoarthritis. *Gait Posture* 2017;57:301–2.
64. Rabin A, Kozol Z. Measures of range of motion and strength among healthy women with differing quality of lower extremity movement during the lateral step-down test. *J Orthop Sports Phys Ther* 2010;40:792–800.
65. Rabin A, Kozol Z, Moran U, *et al*. Factors associated with visually assessed quality of movement during a lateral step-down test among individuals with patellofemoral pain. *J Orthop Sports Phys Ther* 2014;44:937–46.
66. Cornell DJ, Ebersole KT. INTRA-RATER test-retest reliability and response stability of the FUSIONETICS™ movement efficiency test. *Int J Sports Phys Ther* 2018;13:618–32.
67. Maclachlan L, White SG, Reid D. Observer rating versus three-dimensional motion analysis of lower extremity kinematics during functional screening tests: a systematic review. *Int J Sports Phys Ther* 2015;10:482–92.
68. Cronström A, Creaby MW, Nae J, *et al*. Modifiable factors associated with knee abduction during weight-bearing activities: a systematic review and meta-analysis. *Sports Med* 2016;46:1647–62.
69. Streiner DL, Norman GR, Cairney J. *Health measurement scales : a practical guide to their development and use*. Oxford University Press: Oxford, 2015.
70. Fjellner A, Bexander C, Faleij R, *et al*. Interexaminer reliability in physical examination of the cervical spine. *J Manipulative Physiol Ther* 1999;22:511–6.
71. Jonsson A, Rasmussen-Barr E. Intra- and inter-rater reliability of movement and palpation tests in patients with neck pain: a systematic review. *Physiother Theory Pract* 2018;34:165–80.
72. Pool JJ, Hoving JL, de Vet HC, *et al*. The interexaminer reproducibility of physical examination of the cervical spine. *J Manipulative Physiol Ther* 2004;27:84–90.
73. Stochkendahl MJ, Christensen HW, Hartvigsen J, *et al*. Manual examination of the spine: a systematic critical literature review of reproducibility. *J Manipulative Physiol Ther* 2006;29:475–85.
74. Higgins JPT, Thompson SG, Deeks JJ. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.