BMC Sports Science, Medicine and Rehabilitation

RESEARCH Open Access

Visual assessment of movement quality: a study on intra- and interrater reliability of a multi-segmental single leg squat test

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

Abstract

Background: The Single Leg Squat test (SLS) is a common tool used in clinical examination to set and evaluate rehabilitation goals, but there is not one established SLS test used in the clinic. Based on previous scientific findings on the reliability of the SLS test and with a methodological rigorous setup, the aim of the present study was to investigate the intra- and interrater reliability of a standardised multi-segmental SLS test.

Methods: We performed a study of measurement properties to investigate the intra- and interrater reliability of a standardised multi-segmental SLS test including the assessment of the foot, knee, pelvis, and trunk. Novice and experienced physiotherapists rated 65 video recorded SLS tests from 34 test persons. We followed the Quality Appraisal for Reliability Studies checklist.

Results: Regardless of the raters experience, the interrater reliability varied between "moderate" for the knee variable (κ = 0.41, 95% CI 0.10–0.72) and "almost perfect" for the foot (κ = 1.00, 95% CI 1.00–1.00). The intrarater reliability varied between "slight" (pelvic variable; κ = 0.17, 95% CI -0.22-0.55) to "almost perfect" (foot variable; κ = 1.00, 95% CI 1.00–1.00; trunk variable; κ = 0.82, 95% CI 0.66–0.97). A generalised kappa coefficient including the values from all raters and segments reached "moderate" interrater reliability (κ = 0.52, 95% CI 0.43–0.61), the corresponding value for the intrarater reliability reached "almost perfect" (κ = 0.82, 95% CI 0.77–0.86).

Conclusions: The present study shows a "moderate" interrater reliability and an "almost perfect" intrarater reliability for the variable all segments regardless of the raters experience. Thus, we conclude that the proposed standardised multi-segmental SLS test is reliable enough to be used in an active population.

Keywords: Single leg squat, Visual assessment, Movement quality, Reliability, Functional tests, Kappa, Reproducibility

Background

In the clinical setting, visual assessment of movement quality is one of the most commonly used methods to examine patients, and to evaluate and target rehabilitation goals. The term movement quality is often used in relation to the visual assessment of asymmetries, compensatory movements, impairments, and efficiency during a functional movement [1, 2]. Movement quality is described as an independent attribute, and unlike quantitative measures such as power and strength, movement quality aims to capture other important aspects of the movement [1, 3, 4]. This is recommended for example in the rehabilitation of anterior cruciate ligament injuries where the assessment of quantitative as well as qualitative aspects are recommended in the decision of a safe return to play [5]. In addition, observation of the alignment of body segments and the maintenance of a correct posture is often included in the assessment

* Correspondence: John.Ressman@ki.se
[1]Department of Neurobiology, Karolinska Institutet, Care Sciences and Society, Division of Physiotherapy, Alfred Nobels Allé 23, 141 83 Huddinge, Sweden
Full list of author information is available at the end of the article

of movement quality [4, 6, 7], and malalignments of the lower extremity segments are often seen in knee injuries and other overuse injuries [8–13].

The Single Leg Squat test (SLS) is a functional movement test widely used in clinical settings to visually assess movement quality of the lower extremity and is proposed to have biomechanical and neuromuscular similarities to a wide range of athletic movements as it simulates common athletic positions such as cutting, jumping, and landing [14, 15]. It is also commonly included in various screening and test batteries used in sports medicine [16–19]. The SLS test has been named, described, performed, and assessed in many different ways, meaning that there is not one established SLS test [20]. Reported performance differs in many aspects of the test, such as depth of the squat, position of the arms, support and the position of the non-weight bearing leg (in front, behind or below the trunk) [18, 21–26]. In addition to the SLS test, the Forward Step Down (FSD) and Lateral Step Down (LSD), are tests performed on a 15–25 cm high box but otherwise performed and assessed in the same manner as the SLS test [23, 27]. Although the movement pattern during the descendent phase of a SLS, FSD or LSD are the same [28, 29], different kinematic and kinetic have been reported between the SLS tests [28], the SLS test and FSD [29] and in addition between men and women [30, 31]. One important aspect is the position of the non-weight bearing leg where the behind position seems to have the most kinematic differences from the front or below position [28].

The SLS test has been reported to be reliable and valid in clinical and research settings for an asymptomatic healthy population when assessing the knee in relation to the foot [20, 21, 32, 33]. In addition, a multi-segmental approach was recently proposed to be feasible and reliable, preferably with a two- or three-point rating scale [20]. The reliability of the SLS test has previously been explored by either rating video recordings of the test or by rating the performance live.

A reference method for measuring movements are 3-dimensional (3D) analysis systems or 2-dimensional (2D) techniques, however not accessible for all clinicians, and is in addition time consuming, impractical, and not applicable in a larger population [34]. Thus, it is important to further develop movement quality tests used in the clinic regarding their measurement properties.

It would be desirable to evolve a less complex and well-defined SLS test, which is easy to use regardless of the examiner's education or clinical experience. The interpretation of the SLS test should in addition comprise a distinct protocol on how to rate the movement. We propose a SLS test, taking the visual assessment of the kinetic chain from the foot to the trunk into consideration; a multi-segmental approach which might give the clinician further information in the clinical assessment and targeted rehabilitation [20, 33]. In the proposed test, we have included an item considering the position of the foot, in contrast to most other SLS tests, as we believe that the foot position affects the alignment of the kinetic chain. The proposed SLS test is based on the findings from two previous meta-analyses on the validity and reliability of visually assessed ratings on the lower extremity [32, 33], and in addition a recent meta-analysis on the intra- and interrater reliability of the SLS test [20]. Recent studies on the reliability of the SLS test have reported poor methodological quality, thus further studies with more robust methodological standardisation are warranted [20]. Based on previous scientific findings on the reliability of the SLS test and a robust methodological standardisation, the aim of the present study was to investigate the intra- and interrater reliability of a standardised multi-segmental SLS test.

Methods

Study design

This study investigated the intra- and interrater reliability of video-recorded SLS tests and followed the Quality Appraisal for Reliability Studies checklist (QAREL) [35] which can be found in Additional file 1.

Subjects

Thirty-seven healthy persons (27 women, 10 men) aged 34 (±12) years were recruited via verbal announcements and informational posters at the Karolinska Institutet in Stockholm. Inclusion criteria were men and women, aged 18 to 65. Exclusion criteria were an ongoing musculoskeletal injury in the lower extremity, a history of serious knee disorder (ligament- or meniscal rupture and knee replacement), a neurological disease, or a visual deficiency that could not be corrected with eyeglasses. A written informed consent to agree to participate in the study was obtained for all individual subjects. The study was approved by the Regional Ethical Review Board in Stockholm: Ethical approval Dnr: 2016/595–31 with amendment Dnr 2017/318–32 and the Karolinska Institute to which the ethical approval belongs.

Data collection

Before performing the SLS test, all test persons filled in a questionnaire concerning demographic and background data. The tests were performed in the movement laboratory of the Karolinska Institutet during 21 March and 11 May 2017, and administrated by two of the authors (JR and WG). The SLS tests was recorded in the frontal and the sagittal plane with two orthogonally placed digital video cameras (Axis Communications 210A) at three metres' distance. The cameras were

placed so that the whole body was visible, with a brown even background.

The SLS test

The test persons were first verbally instructed on how to perform the SLS test by one of the test leaders (JR) and was then allowed to practice the test for three times. When performing the test, the test person followed a pre-recorded video clip with precise verbal instructions on how to perform the test (Additional file 2). All participants were instructed to wear tight shorts/tights, a gym/sports top, T-shirt, or a vest.

The test person was instructed by the pre-recorded video to perform the SLS test with the arms folded across the chest, the non-weight bearing leg flexed so the foot was pointing backwards and the knee pointing straight down to the floor, see Fig. 1. The instruction was to position the weight bearing leg along a sagittal placed sticky tape on the floor, so that the toes pointed straight ahead, and the inside of the foot was parallel to the sticky tape. If the test person could not accomplish this, the foot could be placed in such a way that felt comfortable. The test was performed for both right and left leg and started always with the left leg. The test person was instructed by the pre-recorded video to squat down three times in a controlled manner and with the instruction to go as deep as possible without lifting the heel from the ground or flexing the upper body too much. No additional instructions on how to perform the test was given. All video recordings were scrutinised for quality and "additional ques" such as tattoos, surgical scars or other identifying features which could inflate the reliability, furthermore, no reference standard was available for this material [35].

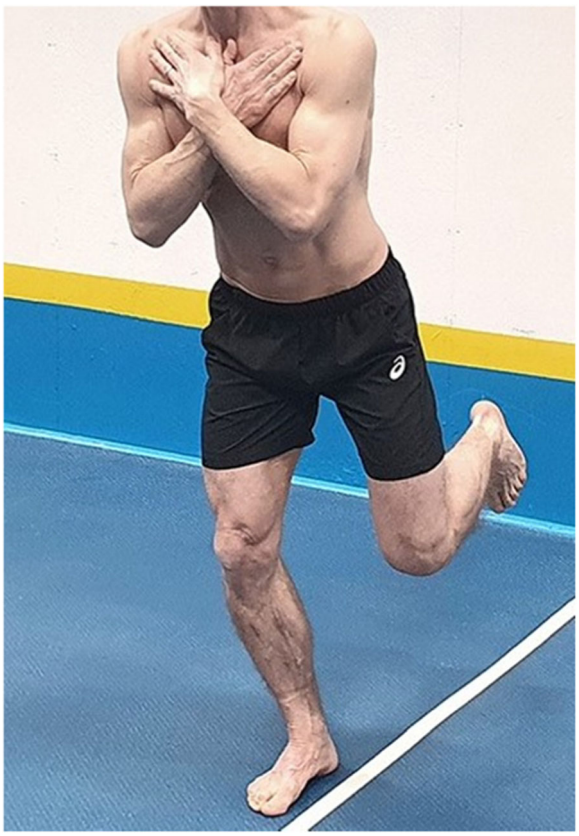

Fig. 1 A correct performance of the Single Leg Squat test

Rating procedure

Raters

Four physiotherapists were included to assess and rate the video recordings: two experienced and two novices. The experienced raters (1 and 2) had more than 20 years of work experience and the novice pair (3 and 4) had about 4 y. The experienced raters worked at a sports medicine clinic where they used specific movement quality tests at a daily basis [17]. The novice raters had no such previous experience in assessing movement quality and had mostly worked in primary health care.

Ten video recordings of the SLS test, along with written instructions on how to rate and assess the tests, were sent to the raters individually. After one week, one of the authors (JR) held a two-hour educational session with all raters. At this session, the ratings of the 10 video recordings were first discussed to reach a consensus on how to rate the test. This was followed by the individual assessment of 10 additional recordings which were then discussed to achieve a consensus on how to assess the SLS test according to the described criteria. Following the educational session, the four raters received 65 new video recordings of the SLS test to assess individually at their own computers for the study purpose. For intrarater reliability, the raters were sent the same video recording after an adequate wash out period of 10 to 14 days [36]. To minimise bias, the order of the videos in the second assessment was randomised with a web-based research randomiser [37]. On both assessment occasions, the raters were instructed to watch each recording no more than two times without any pausing or slow motion. The use of a ruler or any other tool was not allowed. The raters were in addition blinded to each other, their own ratings, and the test persons demographic such as age, activity level and previous injury.

Rating criteria

The rating criteria for the SLS test are described in Table 1. The raters were instructed to observe the video recordings and assess movement deviations from the vertical alignment of the body segments: foot, knee,

Table 1 Rating criteria of the Single Leg Squat test

Observed segments	Correct movement (pass = 0 point)	Movement deviation[a] (fail = 1 point)
Foot[b]		
The relationship of the sagittal plane and metatarsale 2.	Os metatarsale 2 is in relation to the sagittal plane placed in a lateral angel of ≤10°	The metatarsale 2 is in relation to the sagittal plane placed in a lateral angel that **clearly exceeds 10°**
Knee		
Position of the knee in relation to foot.	The centre of the knee is well aligned over the centre of the foot.	The centre of the knee is **clearly** over or medial to digitorum 1.
Medial/lateral perturbation of the knee.	The movement of the knee is vertical and smooth without any medial/lateral shake.	The movement is jerky and **repeated** medial/lateral shake of the knee is seen.
Pelvis		
Lateral pelvic shift and/or pelvic rotation.	No lateral pelvic shift and/or pelvic rotation are seen.	The pelvic is **clearly** shifted lateral and/or rotated in any direction.
Trunk		
Centre of mass: trunk lean, perturbation and balance.	The trunk is well aligned over the pelvic, hip, knee and foot.	The trunk **clearly** leans in either direction, there is **obvious** trunk sway, loss of balance or movement of the arms.

[a]A movement deviation for a segment (1 point) can only be registered one time during the three squats, i.e., a total score of 0–4 points is possible
[b]The position of the foot should be observed before the test is executed. If the test person cannot place the foot in the correct position, they are allowed to put the feet where they feel comfortable
The rater is only allowed to correct the tested person if they:
1. Flex the upper body as much as the hip, pelvis and groin cannot be observed.
2. If the heel is lifted from the ground and/or if the foot is moved from its starting position.
3. If the test person does not understand the instructions and performs a pistol squat instead of the SLS.

pelvis, and trunk during the three consecutive squats. The instruction for this multi-segmental approach was to assess the performance of all body segments at the same time and in relation to each other. A deviation of one segment, could only be scored once (one point) even if failed in all of the three squats. No deviation (pass) was scored as 0 points. The total score for the multi-segmental SLS test could range from 0 to a maximum of 4 points. If scored with 0, no deviations were seen in any of the body segments in any of the three squats. If scored with 4 points, deviation (fail) was evident for all four body segments during any of the three squats.

Statistical analysis

Intra- and interrater reliability was calculated according to Cohen's kappa statistics together with percentage agreement (PA) and a 95% confidence interval (95% CI) for each separate segment: foot, knee, pelvic and trunk variable [38, 39]. Furthermore, for both intra- and inter-rater reliability a merged kappa coefficient was calculated for each segment together and denoted as the variable "all segments." For interrater reliability where multiple raters were compared, a generalised kappa coefficient presented by Fleiss was used [40, 41].

As the magnitude, and interpretation, of the kappa coefficient can be influenced by factors such as prevalence and bias, both prevalence index (PI) and bias index (BI) were calculated and presented together with the kappa statistics (see Tables 3 and 4 for a mathematical clarification) [39]. The effect that prevalence and bias have on the kappa statistics derives from two paradoxes. The first paradox implies that there will be a prevalence effect when there is a predominance of either positive or negative ratings which could be expressed by the PI. A large PI will present a lower kappa and a small PI will present a higher kappa. The effect of PI on kappa is greater for larger values than smaller values [39, 43]. The second paradox relates to the extent of disagreement by the raters on the proportion of positive or negative findings and could be expressed by the BI. A large BI presents a higher kappa, and a small BI presents a lower kappa. The effect of bias is greater when kappa is small and vice versa [39, 43].

As a further support in the interpretation of kappa, the maximum value of kappa ($kappa_{max}$), that could be obtained for the set of data concerned, was also calculated. It is calculated so that the proportions of positive and negative judgements by each rater (i.e., the marginal totals) are taken as fixed, and the distribution of paired ratings (i.e., the cell frequency in the 2×2 tables denoted commonly as a, b, c and d) is adjusted to represent the greatest possible agreement. This means that the maximum possible agreement for either presence or absence of the disease will be the smallest of the marginal totals in each case [39]. $Kappa_{max}$ serves to estimate the strength of the agreement while maintaining the proportions of positive ratings demonstrated by each rater. It provides a reference value for kappa that maintain the individual raters overall tendency to assess a condition or select a rating within the constraints obliged by the marginal totals [39]. Finally, the kappa statistics were

adjusted for low/high bias and prevalence by calculation of the prevalence-adjusted bias-adjusted kappa (PABAK) [39, 43, 44].

The kappa statistics were interpreted according to Landis and Koch classification of strength of agreement [45]; κ:< 0.00 = poor; κ: 0.00–0.20 = slight; κ: 0.21–0.40 = fair; κ: 0.41–0.60 = moderate; κ: 0.61–0.80 = substantial and κ: 0.81–1.0 = almost perfect. Statistical analysis was performed using STATA version 15.1 with the extension of the "kappaetc" command which handles all kappa presented [42], $kappa_{max}$ was calculated via the web calculator [46]. Furthermore, Microsoft Office Excel version 16 for Windows 10 was used for the calculation of PI and BI.

Results

Due to poor video quality, three of the 37 included subjects were excluded and further three subjects could only be assessed for one leg. Hence, in total 65 video recordings and 34 test persons (24 women, 10 men) were included in the study. The test persons had a mean (±SD) age of 34 (12) years and about 80% of those were physically active two days or more per week. The test persons characteristics, pain, and activity levels are described in Table 2. All data from the inter- and intrarater reliability assessment of the SLS test are presented in Tables 3 and 4.

Interrater reliability

For the experienced raters (rater 1 vs. 2), the interrater reliability varied between a "moderate" agreement for the knee variable (κ = 0.42, 95% CI 0.21–0.64) and "almost perfect" for the foot (κ = 1.00, 95% CI 1.00–1.00). The pelvic variable reached a "moderate" agreement (κ = 0.44, 95% CI 0.22–0.66) and the trunk variable a "substantial" agreement (κ = 0.63, 95% CI 0.40–0.85). For the variable all segments, a "moderate" agreement (κ = 0.57, 95% CI 0.46–0.68) was obtained. The largest difference between the calculation of kappa and $kappa_{max}$ was seen for the knee variable (κ = 0.42 vs. $kappa_{max}$ = 0.73), no greater difference was seen between kappa and PABAK.

For the novice raters (rater 2 vs. 3), the interrater reliability varied between a "moderate" agreement for the knee variable (κ = 0.41, 95% CI 0.10–0.72) and "substantial" for the trunk (κ = 0.68, CI 95% 0.46–0.90). The pelvic variable reached a "moderate" agreement (κ = 0.44, 95% CI 0.12–0.76) and the foot variable a "substantial" agreement (κ = 0.66, 95% CI 0.02–1.00). For the variable all segments, a "moderate" agreement (κ = 0.55, 95% CI 0.40–0.70) was obtained. The largest difference between the calculation of kappa and $kappa_{max}$ was seen for the knee variable (κ = 0.41 vs. $kappa_{max}$ = 0.88). In general, PABAK was slightly higher than the kappa coefficient.

For all raters together (rater 1–4), the variable all segments obtained a generalised kappa coefficient of "moderate" agreement 0.52 (95% CI 0.43–0.61), while PABAK reached "substantial" agreement (0.70, 95% CI 0.65–0.76).

Intrarater reliability

For the experienced raters, the intrarater reliability varied between "substantial" (knee variable; κ = 0.71, 95% CI 0.52–0.89) to "almost perfect" agreement (foot variable; κ = 1.00, 95% CI 1.00–1.00). The pelvic variable reached a "substantial" agreement for rater 2 (κ = 0.74, 95% CI 0.51–0.96) and an "almost perfect" agreement for rater 1 (κ = 0.86, 95% CI 0.73–1.00), the trunk variable reached "almost perfect" agreement for both experienced raters (rater 1: κ = 0.89, 95% CI 0.77–1.00; rater 2: κ = 0.95, 95% CI 0.85–1.00). For the variable all segments an "almost perfect" agreement was obtained for both raters (rater 1: κ = 0.93, 95% CI 0.88–0.98; rater 2: κ = 0.82, CI 95% 0.73–0.9). The largest difference between the calculation of kappa and $kappa_{max}$ was seen for rater 2 and the variables knee (κ = 0.71 vs. $kappa_{max}$ = 0.97) and pelvic (κ = 0.74 vs. $kappa_{max}$ = 0.95). No greater difference was seen between kappa and PABAK.

For the novice raters the intrarater reliability ranged from "slight" agreement (pelvic variable; κ = 0.17, 95% CI -0.22-0.55) to "almost perfect" (trunk variable; κ = 0.82, 95% CI 0.66–0.97).

The foot variable varied between a "moderate" agreement for rater 4 (κ = 0.48, 95% CI -0.16-1.00) and a "substantial" agreement for rater 3 (κ = 0.66, 95% CI 0.02–1.00), the knee variable reached "substantial" for both novice raters (rater3: κ = 0.72, 95% CI 0.48–0.96; rater 4: κ = 0.70, 95% CI 0.47–0.92) and the variable all segments reached "substantial" agreement for both raters (rater 3: κ = 0.62, 95% CI 0.45–0.78; rater 4: κ = 0.75, 95% CI 0.64–0.86). The largest difference between the calculation of kappa and $kappa_{max}$ was seen for rater 3 and the variable pelvic (κ = 0.17 vs. $kappa_{max}$ = 0.88) and for rater

Table 2 Test subjects' characteristics, pain, and activity

	All (*n* = 34)	Women (*n* = 24)	Men (*n* = 10)
Age, year			
Mean (SD)	35 (12)	35 (12)	35 (11)
Height, cm			
Mean (SD)	173 (7)	170 (5)	181 (5)
Weight, kg			
Mean (SD)	72 (13)	66 (7)	86 (14)
Physical active ≥2 days/week*			
% of group (*n*)	79% (27)	83% (20)	70% (7)
Pain in regions other than the lower limb			
% of group (*n*)	27% (9)	25% (6)	30% (3)

*Most common physical activities: running/jogging and weightlifting, but yoga, swimming, power walks and cycling were also reported

Table 3 Interrater reliability for experienced raters with > 20 years of clinical experience and novice rater with ≤4 years of clinical experience

Raters	PA[a]	Kappa[b] (CI 95%)	Kappa$_{max}$[c]	PI[d]	BI[e]	PABAK[f] (CI 95%)
Experienced						
Rater 1 vs. Rater 2						
Foot	1.0	1.00 (1.00–1.00)	1.0	0.91	0	1.00 (1.00–1.00)
Knee	0.71	0.42 (0.21–0.64)	0.73	0.09	−0.14	0.42 (0.19–0.64)
Pelvis	0.77	0.44 (0.22–0.66)	0.52	0.46	− 0.20	0.54 (0.33–0.75)
Trunk	0.86	0.63 (0.40–0.85)	0.71	0.52	−0.11	0.72 (0.55–0.90)
All segments[g]	0.84	0.57 (0.46–0.68)	0.71	0.50	−0.11	0.67 (0.58–0.76)
Novice						
Rater 3 vs. Rater 4						
Foot	0.99	0.66 (0.02–1.00)	0.66	0.95	0.02	0.97 (0.91–1.00)
Knee	0.88	0.41 (0.10–0.72)	0.88	0.69	−0.03	0.69 (0.51–0.87)
Pelvis	0.88	0.44 (0.12–0.76)	0.58	0.75	0.09	0.75 (0.60–0.92)
Trunk	0.89	0.68 (0.46–0.90)	0.68	0.58	0.11	0.79 (0.63–0.94)
All segments[g]	0.90	0.55 (0.40–0.70)	0.79	0.75	0.05	0.80 (0.73–0.87)
All raters	**PA[a]**	**Generalised kappa[h] (CI 95%)**				**PABAK[f] (CI 95%)**
Rater 1–4						
All segments[g]	0.85	0.52 (0.43–0.61)				0.70 (0.65–0.76)

[a]***PA*** Percent agreement

[b]**Kappa:** Cohen's kappa, calculated by; $\kappa = \frac{P_o - P_c}{1 - P_c}$

Where; P_o (observed agreement) $= \frac{a+d}{n}$ and P_c (chance agreement) $= \frac{(\frac{f_1 x g_1}{n}) + (\frac{f_2 x g_2}{n})}{n}$

[c]**Kappa$_{max}$:** Is calculated so that the proportions of positive and negative judgements by each rater (i.e. the marginal totals) are taken as fixed, and the distribution of paired ratings (i.e. the cell frequency a,b,c and d) is adjusted so as to represent the greatest possible agreement. That will say, the maximum possible agreement for either presence or absence of the disease is the smaller of the marginal totals in each case [39]

[d]***PI*:** Prevalence index, calculated by; $PI = \frac{a-d}{n}$

[e]***BI*** Bias index, calculated by; $BI = \frac{b-c}{n}$

[f]***PABAK*:** Prevalence-adjusted bias-adjusted kappa, calculated by; $PABAK = 2P_0 - 1$

[g]**All segments:** Denotes a merged kappa coefficient for the interrater reliability of each of the segments together (foot, knee, pelvis and trunk)

[h]**Generalised kappa:** A generalisation of Scott's pi presented by Fleiss in order to calculate the interrater reliability of multiple raters [40, 42]

4 and the variable foot (κ = 0.48 vs. kappa$_{max}$ = 1.0). These segments also showed a great difference between kappa and PABAK; pelvic (κ =0.17, 95% CI − 0.22-0.55 vs. PABAK = 0.79, 95% CI 0.63–0.94) and foot (κ =0.48, 95% CI − 0.16-1.00 vs. PABAK = 0.94, 95% CI 0.85–1.00).

For the variable all segments, an overall average kappa was calculated for all raters (rater 1–4) which reached "almost perfect" agreement (κ = 0.82, 95% CI 0.77–0.86), no greater difference was seen between kappa and PABAK.

Discussion

The aim of the present study was to investigate the intra- and interrater reliability of a standardised multi-segmental SLS test. All in all, the SLS test showed an acceptable intrarater reliability for all raters and all separate variables (foot-, knee-, pelvis- and trunk). For all variables, the agreement was classified as "moderate" or better than so (κ ≥0.41), except for the pelvic variable for one of the novices raters. Regardless of the raters experience, and for the variable all segments, the SLS test demonstrated a "moderate" interrater reliability and an "almost perfect" intratater reliability.

In general, reliability is considered to depend on several 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 [33, 47]. Compared to our findings a recent meta-analysis on the intra- and interrater reliability of different SLS tests (SLS, FSD and LSD) [20] included 17 studies investigating the reliability of multi-segmental SLS tests. Seven of those reported higher reliability [7, 23, 24, 48–51], and 10 equivalent reliability [17–19, 22, 52–57] compared to our results. The reason for the higher reliability might be due to several factors, including the methodological setup and actual test performance. Our study used a convenient sample of 34 persons, and 65 video recordings, without any categorisation and equal distribution of the performed tests on the video recordings (i.e., good, fair, or poor performance). In

Table 4 Intratater reliability for experienced raters with > 20 years of clinical experience and novice rater with ≤4 years of clinical experience

Raters	PA[a]	Kappa[b] (CI 95%)	$Kappa_{max}$[c]	PI[d]	BI[e]	PABAK[f] (CI 95%)
Experienced						
Rater 1						
Foot	1.0	1.0 (1.00–1.00)	1.0	0.91	0.00	1.00 (1.00–1.00)
Knee	0.99	0.97 (0.91–1.00)	0.97	−0.06	0.02	0.97 (0.91–1.00)
Pelvis	0.94	0.86 (0.73–1.00)	0.86	0.32	−0.06	0.88 (0.76–1.00)
Trunk	0.95	0.89 (0.77–1.00)	0.96	0.40	0.02	0.91 (0.80–1.00)
All segments[g]	0.97	0.93 (0.88–0.98)	0.98	0.39	−0.01	0.94 (0.90–0.98)
Experienced						
Rater 2						
Foot	1.0	1.0 (1.00–1.00)	1.0	0.91	0.00	1.00 (1.00–1.00)
Knee	0.86	0.71 (0.52–0.89)	0.97	0.25	−0.02	0.72 (0.55–0.90)
Pelvis	0.92	0.74 (0.51–0.96)	0.95	0.65	0.02	0.85 (0.71–0.98)
Trunk	0.99	0.95 (0.85–1.00)	0.95	0.62	0.02	0.97 (0.91–1.00)
All segments[g]	0.94	0.82 (0.73–0.91)	0.99	0.60	0.00	0.89 (0.83–0.94)
Novice						
Rater 3						
Foot	0.99	0.66 (0.02–1.00)	0.66	0.95	0.02	0.97 (0.91–1.00)
Knee	0.92	0.72 (0.48–0.96)	0.94	0.68	−0.02	0.85 (0.71–0.98)
Pelvis	0.89	0.17 (−0.22–0.55)	0.88	0.86	−0.02	0.79 (0.63–0.94)
Trunk	0.92	0.69 (0.43–0.95)	0.94	0.71	−0.02	0.85 (0.71–0.98)
All segments[g]	0.93	0.62 (0.45–0.78)	0.96	0.80	0.01	0.86 (0.80–0.92)
Novice						
Rater 4						
Foot	0.97	0.48 (−0.16–1.00)	1.0	0.94	0.00	0.94 (0.85–1.00)
Knee	0.91	0.70 (0.47–0.92)	0.70	0.63	0.09	0.82 (0.67–0.96)
Pelvis	0.91	0.69 (0.46–0.93)	0.90	0.63	0.03	0.82 (0.67–0.96)
Trunk	0.92	0.82 (0.66–0.97)	0.82	0.40	0.08	0.85 (0.71–0.98)
All segments[g]	0.93	0.75 (0.64–0.86)	0.83	0.65	0.05	0.85 (0.79–0.92)
Rater 1–4	**PA[a]**	**Overall kappa[h] (CI 95%)**	**$Kappa_{max}$[c]**	**PI[d]**	**BI[e]**	**PABAK[f] (CI 95%)**
All segments[g]	0.94	0.82 (0.77–0.86)	0.97	0.61	0.01	0.89 (0.86–0.91)

[a]**PA:** Percent agreement

[b]**Kappa:** Cohens´ kappa, calculated by; $\kappa = \frac{P_o - P_c}{1 - P_c}$

Where; P_o (observed agreement)$= \frac{a+d}{n}$ and P_c (chance agreement) $= \frac{(\frac{f_1 \times g_1}{n}) + (\frac{f_2 \times g_2}{n})}{n}$

[c]**$Kappa_{max}$:** Is calculated so that the proportions of positive and negative judgements by each rater (i.e. the marginal totals) are taken as fixed, and the distribution of paired ratings (i.e. the cell frequency a,b,c and d) is adjusted so as to represent the greatest possible agreement. That will say, the maximum possible agreement for either presence or absence of the disease is the smaller of the marginal totals in each case [39]

[d]***PI*:** Prevalence index, calculated by; $PI = \frac{a-d}{n}$

[e]***BI*:** Bias index, calculated by; $BI = \frac{b-c}{n}$

[f]**PABAK:** Prevalence-adjusted bias-adjusted kappa, calculated by; $PABAK = 2P_0 - 1$

[g]**All segments:** Denotes a merged kappa coefficient for the intrarater reliability of each segments together (foot, knee, pelvis and trunk)

[h]**Overall kappa:** Presents an overall average kappa for the variable all segments for all raters comparing test occasion one and two. Calculated with Cohens'kappa

addition, our raters were instructed to watch the video recordings only twice without any pausing or slow-motion. Crossly et al. [7] and Herman et al. [56] presented "moderate" to "substantial" interrater reliability but used in contrast to our study a consensus panel and six to 15 video recordings, that unlike the other recordings, had been rated with a 100% agreement by the panel at their first rating. Furthermore, McKeown et al. [18] who presented "moderate" interrater reliability allowed their raters' to watch 17 video recordings an unlimited number of times, both in real time and in slow motion. The results of these studies show that the

methodology of a study is affecting the results of reliability to a large extent. We have in our study aimed to resemble a clinical situation and our intention was to evolve a less complex and well-defined multi-segmental SLS test which would be easily used regardless of the examiner's education or clinical experience. The complexity was reduced by using a dichotomous rating scale, not including all possible segments in the kinetic chain, and by taking less movement deviations per segment into account. We used individual training of the raters using 10 video clips and in addition a two-hour educational session to improve the ratings. Seven comparable studies which included both experienced and unexperienced physiotherapists, physiotherapy students and novice athletic therapists showed both better and equivalent reliability than our study but used twice as much (or more) education if not taking the individual training of 10 video clips into account [17, 22, 23, 48, 50, 51, 55]. Thus, it seems that the results from the present multi-segmental SLS test, despite less education, is in accordance with other multi-segmental reliability studies on the SLS test.

It could be discussed if some facilitating utilities assisting the assessment may lead to better reliability. Three comparable studies which showed "substantial" reliability used markers on the floor to indicate the first or second toe, and in addition markers on the tuberosities tibia [23, 50, 55]. It is not really possible to say if their interrater reliability was due to those markers as they also used an extensive education program (4-, 5- and 20-h respectively) and a different methodological setup in comparison to the present study. However, it is interesting to note that Rabin et al. [51, 55] who performed two almost identical studies, except for the population and facilitating utilities, reached "moderate" reliability in the first study [55] and "almost perfect" in the second study [51]. In their second study, they used a vertical pole in addition to the markers, positioned in front of the tested subjects to enhance the visibility of the movements of the lower limb. On the other hand, it might be more likely that the use of the same raters with an additional four-hour education would have made a greater impact on the reliability than the utilities. Our study used a sticky tape placed on the floor with the purpose to mark the sagittal plane when assessing the habitual placement of the foot. It could be so, that the sticky tape facilitated the assessment of the foot but not the knee, which might be reflected by the constant relatively lower kappa statistics for the knee variable.

To our knowledge, so far, no study has investigated the intra- and interrater reliability of the foot position in relation to the sagittal plane. More commonly, the pronation of the foot is considered as a movement deviation and therefore included in the assessment of a SLS test. To provide for the position of the foot, some studies used a sticky tape shaped as a T, or just a verbal instruction to align the foot in the sagittal plane [21, 53, 54, 58], but far from all studies report a standardised foot position. Our study used a standardised foot position which has been described as an alignment of the second metatarsal in relation to the sagittal plane (a lateral angel of ≤10°) [59]. The position of the foot is important as it acts as a specific reference point in the assessment of the knee, but also as an overall reference for the whole kinetic chain. If a test person shows a habitual foot position with a lateral angle ≥10°; it is the authors opinion that the knee in most of these cases will be assessed as a failure. This due to the knee will be positioned medial to the foot or greater toe from the start, even though the movement of the knee might be smooth, vertical, and sagittal aligned. This could also apply to the whole kinetic chain, which could be well aligned over a lateral rotated foot. On the other hand, to force someone into a smaller lateral angel than their habitual foot position might produce movement deviations further up in the chain. This discussion is lacking in the literature and further studies are warranted to investigate the relationship of the foot position and the outcome of a multi-segmental SLS test.

The present study used recorded video clips to observe and assess the performed SLS test. Video recordings were chosen to standardise the testing procedure enabling several raters to assess identical test performances. However, in a clinical situation the therapist most likely will observe and assess the SLS test live meaning that the present method used lowers the tests' ecological validity. As for any test, it is important that the patient understands the instructions of how to perform the test. We therefore recommend that the instructions to perform the present standardised SLS test (Additional file 2) are followed. To assess a SLS test using a multi-segmental approach, all segments are assessed at the same time and in relation to each other. This means that the rater needs to assess the whole kinetic chain at the same time and not one segment at a time. This way of assessing the SLS test has previously been described in studies of the SLS test [6, 7, 17]. In addition, we do not propose a composite score for the SLS test [24, 55] since a total score conceals the information on which segment or segments that have been scored as fail.

Methodological considerations

Three major strengths of the present study are the use of different statistical computations, the methodological standardisation based on the Quality Appraisal for Reliability Studies checklist (QAREL) [35, 60], and that the proposed SLS test was based on findings from previous studies investigating the SLS tests measurement properties [20, 32, 33].

As the magnitude of kappa is influenced by different factors, for example prevalence and bias, a comparison of the strength of kappa across studies with different statistics could be difficult [39, 61]. In this context, kappa$_{max}$ and PABAK acts as a further support in the judgement of the magnitude of an obtained kappa coefficient [39] and enables a robust result in present study. Hence, when taking the prevalence and bias effects acting on the kappa coefficient of the present study in account and considering the particular methodological context in which the study is conducted, we conclude that the proposed multi-segmental SLS test is reliable enough to be used on an active population in the clinical practice. For reliability and validity studies a sample size of at least 50 measures is recommended [61, 62]. The present study used 260 separate measures for each rater (65 video recordings and 4 segments), which could be considered as an appropriate amount of data fulfilling the requirement of at least 50 data points. Even though 3D and 2D studies 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 [20, 33]. The use of video recordings in present study could therefore be considered a strength for the assessment of the intrarater reliability, since the recordings eliminate the normal within-subject variety. On the other hand, a drawback with video recordings is that the authentical patient-clinician interaction is lost. The study population was a convenience sample of both men (29%) and women (71%) with an average age of 34 (±SD 12) years who were relatively active, mostly with running/jogging and weightlifting. This is an appropriate subgroup of subjects where the SLS test could be applicated, increasing the external validity. However, no further generalisations to another population can be made from our findings, and a more equal distribution of men and women would have been preferable. Another limitation of the present study is that no further generalisation across raters or clinicians can be done from our four raters. In contrast to this, Herman et al. [56] included 142 physiotherapists with varying experience and reached equal reliability as present study. On the other hand, as mentioned above, Herman et al. [56] used a methodological setup which might not be comparable with present study. Also, Teyhen et al. [50] used a multi-rater setup, and included 29 doctoral students with less clinical experience, they used an extensive 20-h education program and reach slightly better reliability than present study.

Conclusion

We propose a SLS test, analysed in a study with a rigorously methodological set up, taking the functional aspects of sport-related actions into account, and considering the whole kinetic chain. Regardless of the raters' experience and with a common two-hour education, the present study shows a "moderate" interrater reliability and an "almost perfect" intrarater reliability for the variable all segments. Thus, we conclude that the standardised multi-segmental SLS test is reliable enough to be used in an active population.

Abbreviations

3D: 3-dimensional; 2D: 2-dimensional; 95% CI: 95% confidence interval; BI: Bias index; FSD: Forward Step Down; LSD: Lateral Step Down; PA: Percent agreement; PABAK: Prevalence-adjusted bias-adjusted kappa; PI: Prevalence index; QAREL: Quality Appraisal for Reliability Studies checklist; SLS: Single Leg Squat

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13102-021-00289-x.

Additional file 1:. Quality Appraisal for Reliability Studies checklist (QAREL). Contains a table with the 11 items of Quality Appraisal for Reliability Studies checklist (QAREL), their answers and explanations.

Additional file 2:. Instructions to the performance of the Single Leg Squat test. Contains written instructions to the test leader regarding the test performance and verbal instructions to the tested subjects.

Acknowledgements

The authors wish to thank all subjects for participating in the study, the raters for their interest and commitment and the Swedish Sports Confederation for financial support.

Authors' contributions

All authors participated in the design of the study. JR and WG collected all data. JR conducted the two-hour education and handled all administration around the two assessment occasions. JR wrote the manuscript and computed the statistical analyses, ERB and WG provided feedback on the analyses and all drafts. All authors read and approved the final draft.

Funding

The Swedish Sports Confederation supplied minor financial support for the raters. Open Access funding provided by Karolinska Institute.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to ethical regulation at the Karolinska Institute but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. A written informed consent to agree to participate in the study was obtained for all individual subjects. The study was approved by the Regional Ethical Review Board in Stockholm: Ethical approval Dnr: 2016/595–31 with amendment Dnr 2017/318–32 and the Karolinska Institute to which the ethical approval belongs.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Neurobiology, Karolinska Institutet, Care Sciences and Society, Division of Physiotherapy, Alfred Nobels Allé 23, 141 83 Huddinge,

Sweden. [2]Women's Health and Allied Health Professionals' Theme, Karolinska University Hospital, Solna, Stockholm 171 76, Sweden.

Received: 21 January 2021 Accepted: 17 May 2021
Published online: 08 June 2021

References

1. McGill S, Frost D, Andersen J, Crosby I, Gardiner D. Movement quality and links to measures of fitness in firefighters. Work. 2013;45(3):357–66 https://doi.org/10.3233/wor-121538.
2. Whittaker JL, Booysen N, de la Motte S, Dennett L, Lewis CL, Wilson D, et al. Predicting sport and occupational lower extremity injury risk through movement quality screening: a systematic review. Br J Sports Med. 2017; 51(7):580–5 https://doi.org/10.1136/bjsports-2016-096760.
3. Frost D, Andersen J, Lam T, Finlay T, Darby K, McGill S. The relationship between general measures of fitness, passive range of motion and whole-body movement quality. Ergonomics. 2013;56(4):637–49 https://doi.org/10.1080/00140139.2011.620177.
4. McCunn R, Aus der Funten K, Fullagar HH, McKeown I, Meyer T. Reliability and association with injury of movement screens: a critical review. Sports Med. 2016;46(6):763–81 https://doi.org/10.1007/s40279-015-0453-1.
5. van Melick N, van Cingel RE, Brooijmans F, et al. Evidence-based clinical practice update: practice guidelines for anterior cruciate ligament rehabilitation based on a systematic review and multidisciplinary consensus. Br J Sports Med. 2016;50(24):1506–15 https://doi.org/10.1136/bjsports-2015-095898.
6. Chmielewski TL, Hodges MJ, Horodyski M, Bishop MD, Conrad BP, Tillman SM. 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(3):122–9 https://doi.org/10.2519/jospt.2007.2457.
7. Crossley KM, Zhang WJ, Schache AG, Bryant A, Cowan SM. Performance on the single-leg squat task indicates hip abductor muscle function. Am J Sports Med. 2011;39(4):866–73 https://doi.org/10.1177/0363546510395456.
8. Aderem J, Louw QA. Biomechanical risk factors associated with iliotibial band syndrome in runners: a systematic review. BMC Musculoskelet Disord. 2015;16(1):356. https://doi.org/10.1186/s12891-015-0808-7.
9. Botha N, Warner M, Gimpel M, Mottram S, Comerford M, Stokes M. Movement patterns during a small knee bend test in academy footballers with femoroacetabular impingement (FAI). Health Sci Working Papers. 2014; 1(10):1–24.
10. 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(2):59–66 https://doi.org/10.2519/jospt.2010.3024.
11. Jimenez-Del-Barrio S, Mingo-Gomez MT, Estebanez-de-Miguel E, Saiz-Cantero E, Del-Salvador-Miguelez AI, Ceballos-Laita L. Adaptations in pelvis, hip and knee kinematics during gait and muscle extensibility in low back pain patients: a cross-sectional study. J Back Musculoskelet Rehabil. 2020; 33(1):49–56 https://doi.org/10.3233/bmr-191528.
12. Shamsi MB, Sarrafzadeh J, Jamshidi A. Comparing core stability and traditional trunk exercise on chronic low back pain patients using three functional lumbopelvic stability tests. Physiother Theory Pract. 2015;31(2): 89–98 https://doi.org/10.3109/09593985.2014.959144.
13. Weiss K, Whatman C. Biomechanics associated with patellofemoral pain and ACL injuries in sports. Sports Med. 2015;45(9):1325–37 https://doi.org/10.1007/s40279-015-0353-4.
14. Alenezi F, Herrington L, Jones P, Jones R. The reliability of biomechanical variables collected during single leg squat and landing tasks. J Electromyogr Kinesiol. 2014;24(5):718–21 https://doi.org/10.1016/j.jelekin.2014.07.007.
15. Zeller BL, McCrory JL, Kibler WB, Uhl TL. Differences in kinematics and electromyographic activity between men and women during the single-legged squat. Am J Sports Med. 2003;31(3):449–56 https://doi.org/10.1177/03635465030310032101.
16. Trulsson A, Garwicz M, Ageberg E. Postural orientation in subjects with anterior cruciate ligament injury: development and first evaluation of a new observational test battery. Knee Surg Sports Traumatol Arthrosc. 2010;18(6): 814–23 https://doi.org/10.1007/s00167-009-0959-x.
17. Frohm A, Heijne A, Kowalski J, Svensson P, Myklebust G. A nine-test screening battery for athletes: a reliability study. Scand J Med Sci Sports. 2012;22(3):306–15 https://doi.org/10.1111/j.1600-0838.2010.01267.x.
18. McKeown I, Taylor-McKeown K, Woods C, Ball N. Athletic ability assessment: a movement assessment protocol for athletes. Int J Sports Phys Ther. 2014; 9(7):862–73.
19. Nae J, Creaby MW, Nilsson G, Crossley KM, Ageberg E. 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(11):863–73 .https://doi.org/10.2519/jospt.2017.7270.
20. Ressman J, Grooten WJA, Rasmussen BE. 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 Exerc Med. 2019;5(1):e000541 https://doi.org/10.1136/bmjsem-2019-000541.
21. Ageberg E, Bennell KL, Hunt MA, Simic M, Roos EM, Creaby MW. Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. BMC Musculoskelet Disord. 2010;11(1):265. https://doi.org/10.1186/1471-2474-11-265.
22. Kennedy MD, Burrows L, Parent E. Intrarater and interrater reliability of the single-leg squat test. Athletic Ther Tod. 2010;15(6):32–6.
23. Park KM, Cynn HS, Choung SD. Musculoskeletal predictors of movement quality for the forward step-down test in asymptomatic women. J Orthop Sports Phys Ther. 2013;43(7):504–10 https://doi.org/10.2519/jospt.2013.4073.
24. Piva SR, Fitzgerald K, Irrgang JJ, Jones S, Hando BR, Browder DA, et al. Reliability of measures of impairments associated with patellofemoral pain syndrome. BMC Musculoskelet Disord. 2006;7(1):33. https://doi.org/10.1186/1471-2474-7-33.
25. Stensrud S, Myklebust G, Kristianslund E, Bahr R, Krosshaug T. 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(7):589–95 https://doi.org/10.1136/bjsm.2010.078287.
26. 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(1):207. https://doi.org/10.1186/1471-2474-13-207.
27. Weir A, Darby J, Inklaar H, Koes B, Bakker E, Tol JL. Core stability: inter- and intraobserver reliability of 6 clinical tests. Clin J Sport Med. 2010;20(1):34–8 https://doi.org/10.1097/JSM.0b013e3181cae924.
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(2):201–11.
29. Lewis CL, Foch E, Luko MM, Loverro KL, Khuu A. Differences in lower extremity and trunk kinematics between single leg squat and step down tasks. PLoS One. 2015;10(5):e0126258. https://doi.org/10.1371/journal.pone.0126258.
30. Khuu A, Lewis CL. Position of the non-stance leg during the single leg squat affects females and males differently. Hum Mov Sci. 2019;67:102506 https://doi.org/10.1016/j.humov.2019.102506.
31. Weeks BK, Carty CP, Horan SA. Effect of sex and fatigue on single leg squat kinematics in healthy young adults. BMC Musculoskelet Disord. 2015;16(1): 271. https://doi.org/10.1186/s12891-015-0739-3.
32. Nae J, Creaby MW, Cronstrom A, Ageberg E. Measurement properties of visual rating of postural orientation errors of the lower extremity - a systematic review and meta-analysis. Phys Ther Sport. 2017. https://doi.org/10.1016/j.ptsp.2017.04.003;27:52–64.
33. 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(3):210–24 https://doi.org/10.1179/1743288x15y.0000000006.
34. 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(1):7–11. https://doi.org/10.1123/jsr.21.1.7.
35. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010;63(8):854–61 https://doi.org/10.1016/j.jclinepi.2009.10.002.
36. Streiner DL, Norman GR, Cairney J. Health measurement scales : a practical guide to their development and use. Oxford: Oxford University Press; 2015.
37. Urbaniak GC, & Plous, S. (2013). Research randomizer (version 4.0) [computer software]. Retrieved on January 1, 2020; Available from: https://www.randomizer.org/
38. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.

39. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68. https://doi.org/10.1093/ptj/85.3.257.
40. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. https://doi.org/10.1037/h0031619.
41. Cho M, Paik P, Joseph L. Statistical Methods for Rates and Proportions. In: Wiley series in probability and statistics. 3rd ed. US: Wiley-Interscience; 2003.
42. Klein D. Implementing a general framework for assessing interrater agreement in Stata. Stata J. 2018;18(4):871–901. https://doi.org/10.1177/1536867X1801800408.
43. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423–9 https://doi.org/10.1016/0895-4356(93)90018-v.
44. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas. 1981;41(3):687–99 https://doi.org/10.1177/001316448104100307.
45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. https://doi.org/10.2307/2529310.
46. VassarStats: Website for Statistical Computation- Kappa as a Measure of Concordance in Categorical Sorting. https://vassarstats.net/kappa.html. Accessed 1 June 2020.
47. Knudson D. What can professionals qualitatively analyze? Journal of physical education. Recreation Dance. 2000;71(2):19–23. https://doi.org/10.1080/07303084.2000.10605997.
48. Junge T, Balsnes S, Runge L, Juul-Kristensen B, Wedderkopp N. 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(1):203. https://doi.org/10.1186/1471-2474-13-203.
49. Poulsen DR, James CR. Concurrent validity and reliability of clinical evaluation of the single leg squat. Physiother Theory Pract. 2011;27(8):586–94 https://doi.org/10.3109/09593985.2011.552539.
50. 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.
51. Rabin A, Kozol Z, Moran U, Efergan A, Geffen Y, Finestone AS. 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(12):937–46 https://doi.org/10.2519/jospt.2014.5507.
52. Barker-Davies RM, Roberts A, Bennett AN, Fong DTP, Wheeler P, Lewis MP. Single leg squat ratings by clinicians are reliable and predict excessive hip internal rotation moment. Gait Posture. 2018;61:453–8 https://doi.org/10.1016/j.gaitpost.2018.02.016.
53. Cornell DJ, Ebersole KT. Intra-rater test-retest reliability and response stability of the Fusioneticstm movement efficiency test. Int J Sports Phys Ther. 2018;13(4):618–32. https://doi.org/10.26603/ijspt20180618.
54. Kaukinen PT, Arokoski JP, Huber EO, Luomajoki HA. Intertester and intratester reliability of a movement control test battery for patients with knee osteoarthritis and controls. J Musculoskelet Neuro Interact. 2017;17(3):197–208.
55. 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(12):792–800 https://doi.org/10.2519/jospt.2010.3424.
56. Herman G, Nakdimon O, Levinger P, Springer S. 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(3):227–32. https://doi.org/10.1123/jsr.2014-0319.
57. Lenzlinger-Asprion R, Keller N, Meichtry A, Luomajoki H. Intertester and intratester reliability of movement control tests on the hip for patients with hip osteoarthritis. BMC Musculoskelet Disord. 2017;18(1):10. https://doi.org/10.1186/s12891-017-1388-5.
58. Örtqvist M, Mostrom 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(12):2060–6 https://doi.org/10.1007/s00167-011-1542-9.
59. Comerford M, Mottram S. Kinetic control : the management of uncontrolled movement. Chatswood: Elsevier Australia; 2012.
60. Lucas N, Macaskill P, Irwig L, Moran R, Rickards L, Turner R, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QARE L). BMC Med Res Methodol. 2013;13(1):111. https://doi.org/10.1186/1471-2288-13-111.
61. HCWd d V, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine : a practical guide. Cambridge: Cambridge University Press; 2011.
62. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651–7 https://doi.org/10.1007/s11136-011-9960-1.

Publisher's Note