ARTICLES

Reliability and Validity of a Weight-Bearing Measure of Ankle Dorsiflexion Range of Motion

Martin D. Chisholm, BSc, BSc(PT), MSc(PT), FCAMPT;[*] *Trevor B. Birmingham, PT, PhD;*[†‡]
Janet Brown, BSc(PT), MEd;[†] *Joy MacDermid, BSc, BSc(PT), MSc, PhD;*[§¶]
Bert M. Chesworth, BA, BSc(PT), MClSc(PT), PhD[†**]

ABSTRACT

Purpose: To examine reliability and validity of the Lunge Test (LT) of dorsiflexion range of motion and determine the impact of different approaches to obtain a score on these parameters. ***Methods:*** Fifty-three patients with ankle injury/dysfunction provided initial assessment data for cross-sectional convergent and known-groups validity analysis with the Pearson coefficient (r) and paired t-test, respectively; data after 4–8 weeks of treatment for longitudinal validity analysis with coefficient r; and data 3 days later for test–retest reliability using the intra-class correlation coefficient (ICC) and minimal detectable change (MDC). LT scores were determined for the affected leg only (LT_{Aff}) and for the difference between the two limbs (LT_{Diff}). Two strategies were used to calculate LT scores: a single series and the mean of three series of lunges. LTs were correlated with the Lower Extremity Functional Scale and Global Foot and Ankle Scale. ***Results:*** Reliability coefficients were high (ICC = 0.93–0.99). The MDC = 1.0/1.5 cm, LT_{Aff}/LT_{Diff}, respectively. Cross-sectional validity was confirmed for LT_{Diff} ($r = -0.40$ to -0.50). Between-limb differences ($p < 0.05$) supported known-groups validity. Longitudinal validity was supported for both LT change scores ($r = 0.39$–0.63). The number of series of lunges used did not impact results. ***Conclusions:*** A single series of lunges produces a reliable LT score. From a validity perspective, clinicians should use LT_{Diff} on initial assessment and either LT to assess change.

Key Words: reproducibility of results; ankle; range of motion, articular; weight-bearing.

RÉSUMÉ

Objectif : Vérifier la fiabilité et la validité du test fonctionnel de flexion du genou vers l'avant (*lunge test*, LT) pour vérifier l'amplitude de la flexion dorsale du pied et évaluer les effets de diverses approches visant à obtenir une cote plus élevée pour ces paramètres. ***Méthode :*** Un échantillon de 53 patients avec blessure ou dysfonction de la cheville a permis de recueillir des données initiales pour procéder à une analyse convergente croisée et des groupes connus en vue d'évaluer la validité des données à l'aide du coefficient de Pearson (r) et du test de t, respectivement; des données après 4 à 8 semaines de traitement pour l'analyse de la validité longitudinale avec le coefficient r, et des données 3 ans plus tard pour la fiabilité test retest avec coefficient de corrélation intraclasse (CCI) et changement minimal détectable (CMD). Les pointages LT ont été établis pour la jambe touchée uniquement (LT_{Aff}) et pour les différences entre les deux membres (LT_{Diff}). Deux stratégies ont été utilisées pour calculer les pointages LT: une série simple et la moyenne de trois séries de flexions. Les LT ont ensuite été corrélés avec les échelles fonctionnelles pour les membres inférieurs, le pied dans son ensemble et la cheville. ***Résultats :*** Les coefficients de fiabilité sont élevés (CCI = 0,93–0,99). Le CMD = 1,0/1,5 cm pour le LT_{Aff}/LT_{Diff}, respectivement. La validité transversale a été confirmée pour LT_{Diff} ($r = -0{,}40$ à $-0{,}50$). Les différences entre les deux membres ($p < 0{,}05$) appuyaient la fiabilité des groupes connus. La validité longitudinale était appuyée pour les deux pointages de changement LT ($r = 0{,}39$–$0{,}63$). Le nombre de séries de flexions utilisées n'a pas eu d'effet sur les résultats. ***Conclusions :*** Une seule série de flexions permet d'obtenir un pointage à LT fiable. Du point de vue de la validité, les cliniciens devraient utiliser le LT_{Diff} pour l'évaluation initiale, et l'un ou l'autre des LT pour évaluer les changements.

Ankle dorsiflexion occurs naturally during many lower-extremity tasks. Reduced ankle dorsiflexion range of motion (DF-ROM) is common in many orthopaedic conditions that confront physiotherapists, including ankle fractures[1] and sprains.[2] Clinicians pay attention to the arthro- and osteokinematics of the ankle during weight-bearing dorsiflexion. Normally, during this movement, the tibia moves forward over the foot as the tibial plafond glides anteriorly on the talar dome.[3] When this accessory glide is limited—for example, due to an anteriorly

From the *Fowler Kennedy Sport Medicine Clinic; †School of Physical Therapy; ‡Wolf Orthopaedic Biomechanics Laboratory, Faculty of Health Sciences; §School of Rehabilitation Science, McMaster University, Hamilton; **Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario; ¶Clinical Research Lab, Hand and Upper Limb Centre (HULC), St. Joseph's Health Centre, London, Ont.

Correspondence to: Bert M. Chesworth, Associate Professor, School of Physical Therapy, Elborn College, The University of Western Ontario, 1201 Western Rd., London, ON N6G 1H1; bcheswor@uwo.ca.

Contributors: All authors designed the study, collected the data, and analyzed and interpreted the data; drafted or critically revised the article; and approved the final draft.

Competing Interests: None declared.

Physiotherapy Canada **2012; 64(4);347–355; doi:10.3138/ptc.2011-41**

Figure 1 The Lunge Test
Note: The large toe and centre of the calcaneus contact the tape measure, and the knee touches the tape on the wall. Contact with the ground is monitored by the physiotherapist. The measurement recorded is the distance in millimetres between the large toe and the wall.

positioned talus in people with chronic ankle instability[4]—the resulting decrease in DF-ROM prevents the ankle joint from achieving a close-packed position of bony stability, making it more vulnerable to inversion and internal rotation forces about the ankle.[3] This is believed to increase the risk of repeated injury, as people with functional ankle instability have demonstrated increases in the vertical component of the ground reaction force in the presence of suboptimal ankle joint positioning when landing from a jumping activity.[5] The importance of normalizing the relationship between physiologic and accessory ankle movements in weight bearing is evident in the development and investigation of treatment approaches that target these components to improve weight-bearing DF-ROM and spatiotemporal postural control.[6,7]

Techniques to measure DF-ROM can be grouped into three categories, based on measurement method and body position: visual estimation, goniometric measurement in non-weight-bearing positions, and measurement in a weight-bearing position. Each grouping demonstrates different levels of reliability. Visual estimation has poor measurement qualities and has not been recommended for use in clinical settings.[8,9] Non-weight-bearing measures of DF-ROM have variable reports of intrarater reliability,[9–13] with intra-class correlation coefficient (ICC) values varying from 0.64[9] to 0.97[13] and interrater ICCs as high as 0.87.[10] For weight-bearing measures of DF-ROM, reports of both types of reliability have been uniformly high, with intrarater ICCs from 0.93[14] to 0.99[2,15] and interrater ICCs of 0.98[14] and 0.99.[16]

Bennell[16] introduced the weight-bearing Lunge Test (LT) for quantifying DF-ROM using a simple tape measure secured to the floor (see Figure 1). One key aspect of the testing protocol is its iterative nature: a series of lunges is performed to determine a single numeric value, and then this procedure is repeated three times, so that a mean of three values represents DF-ROM.

One question about this research method is whether it can be translated directly to clinical practice. In the literature, researchers have used up to six[17] series of lunges to generate a mean value for characterizing DF-ROM. Some investigators have measured only the affected limb;[2,15,16] others have used the difference between limbs as the measure of abnormal DF-ROM.[18] Clinicians working under time constraints in busy treatment settings may choose to use a single series of lunges with the affected limb only; we do not know the impact of using a single versus multiple series of lunges with affected versus bilateral limbs on the reliability of the measure.

The validity of the LT has not been examined to the same extent as its reliability. In healthy study participants, ultrasound images have shown that LT values do correlate with gastrocnemius/soleus muscle fascicle lengths and pennation angles.[19] Among patients with an ankle fracture, two studies have demonstrated the predictive validity of affected-limb LT scores on activity limitation.[1,20]

The present study was designed as an initial parameter estimation study of the psychometric properties of the LT in a sample of orthopaedic patients. Our objectives were to examine the reliability and validity of the LT and to determine the impact on these measurement properties of using a single or multiple series of lunges with one or both limbs.

METHODS

Study design

Figure 2 outlines how data were used to achieve the study objectives. Participants were recruited from four outpatient physiotherapy clinics. Data were collected at the initial assessment (Time 1) for convergent and known-groups validity, during the fourth to eighth week of rehabilitation (Time 2) for longitudinal validity, and within 3 days of Time 2 for test–retest reliability (Time 3). The Health Sciences Research Ethics Board of the University of Western Ontario approved this study, and all participants provided written informed consent.

Participants

Inclusion criteria were as follows: age >18 years, attending physiotherapy for post-surgical or non-surgical unilateral ankle dysfunction of musculoskeletal origin, loss of DF-ROM as judged by the treating physiotherapist, able to read/follow instructions in English, willing to attend testing sessions, and able to provide informed consent. Exclusion criteria were inability to successfully complete the LT, contraindication for full weight bearing, or presence of a concomitant neurological disorder or ankle arthrodesis. The target sample size of 35 was based on two test sessions and parameter estimation of

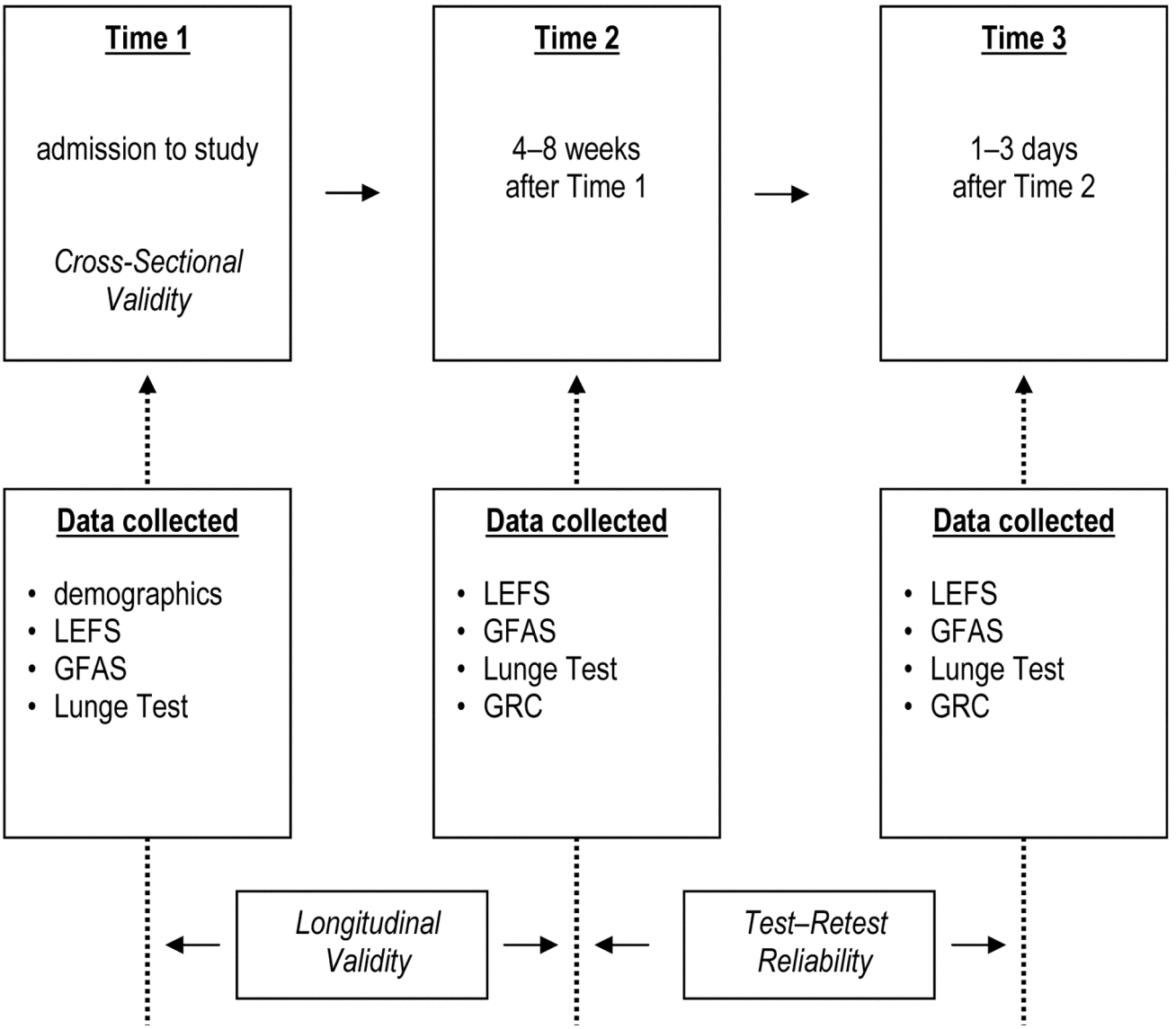

Figure 2 Study timeline and study objectives (in italics).
LEFS = Lower Extremity Functional Scale; GFAS = American Academy of Orthopedic Surgeons Global Foot and Ankle Scale; GRC = Global Rating of Change.

ICC = 0.85 (95% CI width of 0.20)[21] and a 10% loss to follow-up.

Data collection

At Time 1, physiotherapists conducted a typical initial examination. For descriptive purposes, the following components of this assessment were recorded: age (y); sex; height (cm) and weight (kg) for body mass index (BMI) in kg/m^2; mechanism of injury; and date of injury/surgery.

Measures

In addition to the LT, two self-report measures of function were administered at all three time points: the Lower Extremity Functional Scale (LEFS)[22] and the American Academy of Orthopedic Surgeons (AAOS) Global Foot and Ankle Scale (GFAS).[23] At Time 2 and Time 3, the clinician and the participant completed a Global Rating of Change (GRC) score.[24]

Lunge Test of ankle DF-ROM

The LT (see Figure 1) was performed following procedures outlined by Bennell.[16] First, the unaffected foot was placed forward, with the great toe and centre of the heel on the tape measure. With both feet stationary, a controlled forward lunge was performed such that the knee flexed as the participant attempted to touch it to a vertical line marked on the wall with adhesive tape. During this movement, the physiotherapist maintained the foot's alignment on the tape measure secured to the

floor, monitored the heel to ensure contact with the floor, and watched for knee contact with the wall. Pronation or supination of the foot was not controlled. An attempt was considered successful if the participant was able to touch the knee to the wall while maintaining the proper foot alignment and heel contact with the tape on the floor. Upon successfully touching the knee to the wall, the participant moved the foot further from the wall and performed another forward lunge, once again attempting to touch the knee to the wall. The participant was given up to five attempts to achieve the greatest distance between the large toe and the wall. Using the tape measure on the floor, this distance was recorded in millimetres to indicate a single value of DF-ROM. The entire process was then repeated with the affected limb. To obtain three values of DF-ROM for each ankle, the limb testing sequence was unaffected–affected–unaffected–affected–unaffected–affected.

Lower Extremity Functional Scale (LEFS)

The LEFS is a 20-item region-specific self-report questionnaire that asks participants to rate their perceived ability to perform various lower-extremity tasks on a 5-point scale (0 = extreme difficulty or unable to perform activity, 4 = no difficulty). Ratings are summed for a total LEFS score from 0 (very poor function) to 80 (excellent function). First described by Binkley,[22] the LEFS has been shown to have excellent test–retest reliability in people with a variety of musculoskeletal complaints presenting to physiotherapy; good test–retest reliability has been reported in people with ankle sprains.[25]

AAOS Global Foot and Ankle Scale (GFAS)

The GFAS is a 20-item region-specific self-report questionnaire with four domains (pain, function, stiffness and swelling, and giving way) and varying response scales depending on the item. The standardized score (0–100, indicating very poor to excellent function) was used in this study. The GFAS has good internal consistency and acceptable test–retest reliability, and there is evidence to support its validity.[23]

Global Rating of Change scores (GRC)

The GRC uses a 15-point ordinal scale (−7 = very great deal worse, +7 = very great deal better).[26] Norman[27] has questioned the use of retrospective measures of change to recall functional status; to address this concern, Stratford[28] has suggested using both clinician and participant ratings to create an average GRC.

Rater training

The raters used in the study were nine physiotherapists, one physiotherapy assistant, one kinesiologist, and two physiotherapy students. Clinical experience varied from 0 (i.e., the students) to 17 years. A 20-minute training session was used to demonstrate how to conduct the LT and to discuss inclusion and exclusion criteria, obtaining informed consent, administering questionnaires, recording of data, and the study timelines. Raters were not blinded, but they were asked to not review previous findings before performing the testing at Time 2 and Time 3. Periodic visits were conducted during data collection to review procedures. All four clinics were equipped with identical tape measures, and the set-up was reviewed by the primary investigator.

Analysis

Participant characteristics were summarized by means of descriptive statistics. LT scores (mm) were calculated for analyses involving the affected limb (LT_{Aff}) and for the difference between affected and unaffected limbs (LT_{Diff}). For LT_{Diff}, the affected side (LT_{Aff}) was subtracted from the unaffected side (LT_{Unaff}), so a positive value indicates that the unaffected ankle had more DF-ROM than the affected ankle. Values for LT_{Aff} and LT_{Diff} were calculated in two ways: one set of scores used only the first value from the LT protocol, while the other set used the mean of all three measurements obtained from the testing protocol.

Reliability

Test–retest reliability used Time 2 and Time 3 data, because we felt that ankle status should not vary over 1–3 days by the 4- to 8-week mark after beginning treatment. To test this assumption, we compared the means of the LT, LEFS, and GFAS at these two time points using paired *t*-tests.[29] We calculated the $ICC_{2,1}$ with its 95% CI,[29] as well as the standard error of measurement (SEM) and the minimal detectable change at the 90% CI (MDC_{90}).[24,29] The SEM CIs were calculated following Stratford and Goldsmith.[30]

It has been suggested that measures of agreement are more appropriate than reliability measures for tools that will be used to assess clinical change.[31] Therefore, we determined the 95% limits of agreement between the test–retest LT values[32] and calculated the percentage of patients for whom test–retest scores differed by less than two threshold values[31] (5 mm and 10 mm).

We performed two sets of reliability analyses, the first using values from the first measurement obtained from the LT protocol and the second using the mean of all three measurements from the protocol.

Construct validity

Time 1 and Time 2 data were used to examine validity. We used a construct-validation process[33] to examine cross-sectional and longitudinal convergent and known-groups validity, as well as sensitivity to change.

To examine cross-sectional convergent validity, we used the Pearson product–moment correlation coefficient (*r*)[29] to assess the correlation of LT_{Aff} and LT_{Diff} with LEFS and GFAS, both at Time 1 and at Time 2. The hypothesis being tested was that greater DF-ROM should be correlated with better ankle-related function. Since higher LEFS and GFAS scores reflect better function, we expected a

positive correlation with LT_{Aff}, which increases as DF-ROM improves; since an individual with very little side-to-side difference in DF-ROM would have a low LT_{Diff}, we expected a negative correlation with LEFS and GFAS scores. Cross-sectional known-groups validity was examined at Time 1 by comparing LT_{Aff} with LT_{Unaff} using a paired *t*-test. A significant difference between these means would indicate that the LT was able to differentiate between these two known groups.

To examine longitudinal convergent validity, we correlated the change in LT_{Diff} and LT_{Aff} between Time 1 and Time 2 with change scores for the LEFS and the GFAS using the coefficient *r*. The magnitude of these correlations provides information about the extent to which a change in DF-ROM, as measured by the LT, is related to a change in functional ability. We wanted improvement in all measures to be reflected by positive change scores, so that a positive association was reflected by a positive coefficient *r*. We anticipated that the magnitude of LT_{Diff} would decrease as DF-ROM of the affected ankle improved over time; therefore, Time 2 LT_{Diff} scores were subtracted from Time 1 scores, so that a positive value indicated an improvement in DF-ROM. We expected that LT_{Aff} would increase as DF-ROM of the affected ankle improved; therefore, Time 1 LT_{Aff} scores were subtracted from Time 2 scores, so that a positive value would reflect an improvement in DF-ROM. We also anticipated that function would improve over time, and so a participant's score on the LEFS and GFAS would show improvement by increasing in value; therefore, we subtracted participants' LEFS and GFAS Time 1 scores from their Time 2 scores, anticipating a positive correlation.

Sensitivity to change was analyzed using the approach for a heterogeneous sample of individuals, most of whom were expected to change by different amounts.[34] This analysis used the average GRC scores from the participants and physiotherapists. First, we used the $ICC_{3,1}$ with its 2-sided 95% CI[29] to examine the reliability of the average GRC scores between Time 2 and Time 3. Pearson's *r* was then calculated to examine the relationship between the average GRC at Time 2 and the change in LT_{Aff} and LT_{Diff} among participants. A positive correlation was anticipated.

We also calculated the effect size (ES) and standardized response mean (SRM), which, while often considered an inappropriate approach to analyzing sensitivity to change for a heterogeneous sample,[35] are nonetheless frequently reported in the literature. The ES was calculated as the average change between Time 1 and Time 2 divided by the standard deviation of the initial scores,[36] and the SRM as the average change between Time 1 and Time 2 divided by the standard deviation of that change score.[37]

We performed two sets of validity analyses, the first using values from the first measurement obtained from the LT protocol and the second using the mean of all three measurements from the protocol.

RESULTS

Participant characteristics

Study participants were predominantly young, active adults with a mean (SD) age of 34.6 (13.9) years and a BMI of 25.3 (3.0) kg/m^2. As defined by referral diagnosis, the largest group of participants (55%) had an inversion sprain; 15% had an ankle fracture, 11% had tendinopathy, and 7% had an eversion sprain. The rest were referred for osteoarthritis, Achilles tendon repair, surgical stabilization, calf strain, posterior impingement, contusion, or gunshot wound. Other characteristics are shown in Table 1. Of the 53 participants recruited at Time 1, 43 remained at Time 2 (after 4 weeks of rehabilitation), the rest having self-discharged from physiotherapy. The 37 participants who remained at Time 3 were those able to attend the retest session within the 1- to 3-day time window.

Reliability

Test–retest reliability findings are shown in Table 2. Across all approaches for generating a LT score, there was no difference between testing occasions (LT_{Aff}: first test, $t = 1.70$, $df = 36$, $p = 0.10$; mean of 3 tests, $t = 2.06$, $df = 36$, $p = 0.05$. LT_{Diff}: first test, $t = -1.36$, $df = 36$, $p = 0.18$; mean of 3 tests, $t = -1.70$, $df = 36$, $p = 0.10$). All ICC values were >0.90, SEM varied from 4.0 to 5.7 mm, and MDC_{90} varied from 9.4 to 13.3 mm. The GFAS scores were no different between Time 2 and Time 3 ($t = 0.18$, $df = 36$, $p = 0.86$), but LEFS scores at Time 2 differed from those at Time 3 ($t = 2.47$, $df = 36$, $p = 0.019$).

For the agreement parameters in Table 2, across all approaches for calculating an LT score, more than 80% of patients had LT scores that differed by $\leq$10 mm between test occasions. This proportion dropped to less than 70% when the threshold for this difference was $\leq$5 mm.

Validity

For known-groups validity, LT_{Aff} scores were different from LT_{Unaff} scores ($t = -13.71$, $df = 52$, $p < 0.001$). Mean (SD) values for the first measurement from the testing protocol were 56.8 (38.1) mm and 116.2 (35.0) mm, respectively. The corresponding values for the mean of three measurements from the testing protocol (not reported) were similar.

Correlational validity findings are shown in Table 3. For cross-sectional convergent validity of LT_{Aff}, all CIs for Pearson's *r* spanned the null value. For LT_{Diff}, by contrast, no CIs for Pearson's *r* spanned the null value, and the point estimates varied from -0.40 to -0.50. Similar findings (not reported) were found for correlations at Time 2. For longitudinal validity, regardless of the approach to measuring the LT, values of *r* varied from 0.57 to 0.63 for the LEFS and 0.39 to 0.59 for the GFAS, with no CIs spanning the null value. For sensitivity to change, improvement in DF-ROM was associated with average GRC at Time 2. Use of the average GRC was

Table 1 Participant Characteristics by Testing Occasion

Characteristic	Testing occasion; no. (%) of patients*		
	Time 1	Time 2	Time 3
	Initial assessment ($n = 53$)	4–8 wk after Time 1 ($n = 43$)	1–3 d after Time 2 ($n = 37$)
Female sex	24 (45)	20 (47)	17 (46)
Age, y			
18–25	11 (21)	9 (21)	9 (24)
26–35	24 (45)	20 (47)	16 (43)
36–45	7 (13)	3 (7)	3 (8)
46–55	5 (9)	5 (12)	4 (11)
56–65	2 (4)	2 (5)	1 (3)
>65	4 (8)	4 (9)	4 (11)
Affected ankle, right side	26 (49)	23 (53)	20 (54)
Time since injury/surgery			
Acute (≤3 d)	2 (4)	2 (5)	1 (3)
Subacute (4 d to <2 wk)	7 (13)	6 (14)	6 (16)
Early chronic (2–4 wk)	14 (26)	9 (21)	8 (22)
Chronic			
(1–3 mo)	15 (28)	12 (28)	10 (27)
(3–6 mo)	11 (21)	10 (23)	9 (24)
(6–12 mo)	0 (0)	0 (0)	0 (0)
Longstanding (>1 y)	4 (8)	4 (9)	3 (8)
LEFS (0–100); mean (SD)	49.0 (12.4)	62.2 (12.5)	64.9 (11.6)
GFAS (0–100); mean (SD)	68.9 (14.4)	84.1 (11.3)	84.6 (11.0)

*Unless otherwise specified.
LEFS = Lower Extremity Functional Scale (worst–best); GFAS = American Academy of Orthopedic Surgeons Global Foot and Ankle Scale (worst–best).

Table 2 Test–retest Reliability and Agreement Findings by Lunge Test Scoring Strategy ($n = 37$)

Findings	Group; mean (SD), mm*			
	Affected only (LT_{Aff})		Unaffected − affected (LT_{Diff})	
	1st test	Mean of 3 tests	1st test	Mean of 3 tests
LT values				
Test occasion				
Time 2 (test)	85.6 (37.3)	89.4 (37.3)	39.9 (21.6)	39.2 (21.3)
Time 3 (retest)	87.4 (36.6)	91.3 (36.9)	38.1 (21.5)	37.3 (21.5)
Time 3–Time 2	1.8 (6.5)	1.9 (5.5)	−1.8 (8.0)	−2.0 (7.1)
Reliability				
Parameter				
ICC (95% CI)	0.98 (0.98–0.99)	0.99 (0.98–0.99)	0.93 (0.87–0.96)	0.94 (0.89–0.97)
SEM (95% CI)	4.7 (3.8–6.1)	4.0 (3.3–5.2)	5.7 (4.7–7.4)	5.1 (4.2–6.6)
MDC_{90}	10.9	9.4	13.3	11.9
Agreement				
Parameter				
95% limits of agreement	−14.5, 10.9	−12.6, 8.8	−13.9, 17.4	−11.9, 15.8
% ≤5 mm†	65	68	41	65
% ≤10 mm‡	92	86	81	84

*Unless otherwise indicated.
†Percentage of patients with LT values differing ≤5 mm between test occasions.
‡Percentage of patients with LT values differing ≤10 mm between test occasions.
LT = Lunge Test; MDC_{90} = minimal detectable change at the 90% CI.

Table 3 Association between Lunge Test Scores and Self-Report Measures of Function

	Scoring strategy; Pearson correlation coefficient (95% CI)			
	Affected only (LT_{Aff})		Unaffected-affected (LT_{Diff})	
Type of validity	1st test	Mean of 3 tests	1st test	Mean of 3 tests
Cross-sectional*				
LEFS	0.18 (−0.10–0.43)	0.18 (−0.10–0.43)	−0.40 (−0.61 to −0.15)	−0.42 (−0.62 to −0.17)
GFAS	0.20 (−0.08–0.45)	0.20 (−0.08–0.45)	−0.47 (−0.66 to −0.23)	−0.50 (−0.68 to −0.27)
Longitudinal†				
LEFS	0.59 (0.35–0.76)	0.57 (0.33–0.74)	0.59 (0.35–0.76)	0.63 (0.41–0.78)
GFAS	0.41 (0.13–0.63)	0.39 (0.10–0.62)	0.55 (0.30–0.73)	0.59 (0.35–0.75)
Sensitivity to change‡	0.54 (0.29–0.72)	0.56 (0.31–0.74)	0.33 (0.03–0.57)	0.40 (0.11–0.63)

*Correlation between Time 1 values ($n = 53$).
†Correlation between change scores ($n = 43$).
‡Correlation between change scores and Time 2 global ratings of change ($n = 43$).
LT = Lunge Test; LEFS = Lower Extremity Functional Scale; GFAS = American Academy of Orthopaedic Surgeons Global Foot and Ankle Scale.

supported by reliable test–retest ratings: $ICC_{3,1} = 0.91$ (95% CI, 0.84–0.95). Across all approaches for generating an LT score, the ES and SRM varied from 0.70 to 0.73 and 0.99 to 1.00, respectively.

DISCUSSION

This study found good test–retest reliability of the LT in a sample of patients presenting with orthopaedic ankle dysfunction. We have also shown that the LT provided acceptable agreement findings and evidence supporting the validity of the test.

Reliability and agreement

For test–retest reliability, the current findings are similar to published ICC values noted above, which have been consistently high whether participants had no ankle dysfunction[2,15,38] or whether the LT_{Diff} score was used.[18]

Similarly, published values for the SEM and MDC (3 mm and 8 mm, respectively, for intra-observer reliability)[38] are close to those found in the current study. Comparing the MDC_{90} values in Table 2 shows how quicker approaches to quantifying the LT affect the measure's ability to detect true change. If a single series of lunges is used, the MDC_{90} is about 1.5 mm larger than that produced by the more time-consuming approach of taking an average of three tests; if a score is obtained from the affected limb only, the MDC_{90} is about 2.5 mm larger than that produced by measuring both limbs.

In looking further at the MDC_{90} and agreement findings in Table 2, clinicians could conclude that true change in DF-ROM has occurred when LT_{Aff} changes *by about 1 cm*. The high percentage of patients with test–retest values that differed by ≤1 cm supports the use of this value for the MDC_{90} when measuring the affected ankle. When the LT_{Diff} score is the variable of interest, clinicians could conclude that true change in DF-ROM has taken place when the LT_{Diff} score changes *by about 1.5 cm*. Once again, the high percentage of patients with test–retest values that differed by ≤1 cm supports the use of this higher value for the MDC_{90} when measuring both limbs. These MDC values align well with the fact that tape measures typically mark 1 cm and 0.5 cm increments prominently. Future intervention studies should report the proportion of study participants who achieve these MDC values, to strengthen their clinical utility.

Validity

Our validity results agree with previous reports that performance-based measures and self-report functional measures are, at best, moderately correlated ($r < 0.60$).[25,39] For example, Alcock and Stratford[25] found a correlation of 0.36 between non-weight-bearing DF-ROM and LEFS scores for people with ankle sprains. In addition, the cross-sectional correlations between LT_{Aff} and the LEFS in our study are similar to those reported for patients with an ankle fracture after cast removal (95% CIs for $r = 0.07$–0.21 at 6 weeks and 0.05–0.20 at 6 months).[1]

As Table 3 shows, the validity findings suggest a measurement strategy that may minimize clinical time spent measuring DF-ROM. The absence of a cross-sectional relationship between LT_{Aff} and LEFS and GFAS scores, in the presence of a longitudinal relationship between their change scores, suggests that measurement of LT_{Aff} is a valid means of documenting change. The moderate cross-sectional and longitudinal correlations between LT_{Diff} and LEFS and GFAS scores support the cross-sectional and longitudinal convergent validity of the LT_{Diff}, which suggests that clinicians should measure LT_{Diff} at the initial assessment if the goal is to provide a valid measure of the current status of ankle mobility relative to the uninvolved limb. When clinicians seek to document within-limb change over time, longitudinal validity findings support the use of the less time-consuming LT_{Aff}.

Our study has several limitations. First, the findings are not generalizable outside of the active adult popula-

tion presenting with orthopaedic ankle conditions. Second, only individuals who could perform the LT as described by Bennell[16] were included; patients with weight-bearing restrictions and those unable to perform the LT are therefore not represented in our results. Third, statements about the validity of the test are made from the perspective that this portion of the study, as a parameter-estimation study, was intended to begin the process of examining LT validity. Fourth, the actual time taken to perform the various LT scoring methods was not measured; future study is warranted to determine the relationship between LT scoring strategies and their completion time in clinical settings.

CONCLUSIONS

Our study has shown that the LT has sound reliability and agreement qualities. Known-groups validity of the test is supported. Cross-sectional convergent validity is supported when both limbs are measured. Longitudinal convergent validity is supported when one or both limbs are measured. A single series of up to five lunges can be used to obtain a reliable LT score.

KEY MESSAGES

What is already known on this topic

The Lunge Test (LT) to measure weight-bearing ankle dorsiflexion is a reliable test; however, there is limited evidence on its validity. It is not known whether a quicker testing protocol than the one described for research settings is valid and reliable.

What this study adds

A single series of up to five lunges can be used to generate a reliable LT score. From a validity perspective, our findings suggest that a LT difference score between a participant's ankles should be measured at the initial assessment, but measures of the affected limb alone can be made for the purpose of documenting clinical change.

REFERENCES

1. Hancock MJ, Herbert RD, Stewart M. Prediction of outcome after ankle fracture. J Orthop Sports Phys Ther. 2005;35(12):786–92. Medline:16848099
2. Collins N, Teys P, Vicenzino B. The initial effects of a Mulligan's mobilization with movement technique on dorsiflexion and pain in subacute ankle sprains. Man Ther. 2004;9(2):77–82. http://dx.doi.org/10.1016/S1356-689X(03)00101-2. Medline:15040966
3. Hertel J. Functional anatomy, pathomechanics, and pathophysiology of lateral ankle instability. J Athl Train. 2002;37(4):364–75. Medline:12937557
4. Wikstrom EA, Hubbard TJ. Talar positional fault in persons with chronic ankle instability. Arch Phys Med Rehabil. 2010;91(8):1267–71. http://dx.doi.org/10.1016/j.apmr.2010.04.022. Medline:20684909
5. Delahunt E, Monaghan K, Caulfield B. Changes in lower limb kinematics, kinetics, and muscle activity in subjects with functional instability of the ankle joint during a single leg drop jump. J Orthop Res. 2006;24(10):1991–2000. http://dx.doi.org/10.1002/jor.20235. Medline:16894592
6. Hoch MC, McKeon PO. The effectiveness of mobilization with movement at improving dorsiflexion after ankle sprain. J Sport Rehabil. 2010;19(2):226–32. Medline:20543222
7. Hoch MC, McKeon PO. Joint mobilization improves spatiotemporal postural control and range of motion in those with chronic ankle instability. J Orthop Res. 2011;29(3):326–32. http://dx.doi.org/10.1002/jor.21256. Medline:20886654
8. Croxford P, Jones K, Barker K. Inter-tester comparison between visual estimation and goniometric measurement of ankle dorsiflexion. Physiother Theory Pract. 1998;14(2):107–13. http://dx.doi.org/10.3109/09593989809057153
9. Youdas JW, Bogard CL, Suman VJ. Reliability of goniometric measurements and visual estimates of ankle joint active range of motion obtained in a clinical setting. Arch Phys Med Rehabil. 1993;74(10):1113–8. http://dx.doi.org/10.1016/0003-9993(93)90071-H. Medline:8215866
10. Diamond JE, Mueller MJ, Delitto A, et al. Reliability of a diabetic foot evaluation. Phys Ther. 1989;69(10):797–802. Medline:2780806
11. Elveru RA, Rothstein JM, Lamb RL. Goniometric reliability in a clinical setting. Subtalar and ankle joint measurements. Phys Ther. 1988;68(5):672–7. Medline:3362980
12. Jonson SR, Gross MT. Intraexaminer reliability, interexaminer reliability, and mean values for nine lower extremity skeletal measures in healthy naval midshipmen. J Orthop Sports Phys Ther. 1997;25(4):253–63. Medline:9083944
13. Van Gheluwe B, Kirby KA, Roosen P, et al. Reliability and accuracy of biomechanical measurements of the lower extremities. J Am Podiatr Med Assoc. 2002;92(6):317–26. Medline:12070231
14. Venturini C, Ituassu NT, Teixeira LM, et al. Intrarater and interrater reliability of two methods for measuring the active range of motion for ankle dorsiflexion in healthy subjects. Rev Bras Fisioter. 2006;10:377–81.
15. Vicenzino B, Prangley I, Martin D. The initial effect of two Mulligan mobilisation with movement treatment techniques on ankle dorsiflexion. In: A sports medicine odyssey—challenges, controversies and change: proceedings of the Australian Conference of Science and Medicine in Sport; 2001 Oct 23–27; Perth, Australia. Available from: http://fulltext.ausport.gov.au/fulltext/2001/acsms
16. Bennell KL, Talbot RC, Wajswelner H, et al. Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. Aust J Physiother. 1998;44(3):175–80. Medline:11676731
17. Jones R, Carter J, Moore P, et al. A study to determine the reliability of an ankle dorsiflexion weight-bearing device. Physiotherapy. 2005;91(4):242–9. http://dx.doi.org/10.1016/j.physio.2005.04.005
18. Vicenzino B, Branjerdporn M, Teys P, et al. The initial effects in posterior talar glide and dorsiflexion after mobilization with movement in individuals with recurrent ankle sprain. J Orthop Sport Phys. 2006;36(7):464–71. http://dx.doi.org/10.2519/jospt.2006.2265
19. Hallet G, McEwan I, Thom J. Validation of the dorsiflexion lunge test. In: Proceedings of the XIV International Congress on Sports Rehabilitation and Traumatology; 2005 Apr 9–10; Bologna, Italy.
20. Lin CW, Moseley AM, Herbert RD, et al. Pain and dorsiflexion range of motion predict short- and medium-term activity limitation in people receiving physiotherapy intervention after ankle fracture: an observational study. Aust J Physiother. 2009;55(1):31–7. http://dx.doi.org/10.1016/S0004-9514(09)70058-3. Medline:19400023
21. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21(9):1331–5. http://dx.doi.org/10.1002/sim.1108. Medline:12111881
22. Binkley JM, Stratford PW, Lott SA, et al, and the North American Orthopaedic Rehabilitation Research Network. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. Phys Ther. 1999;79(4):371–83. Medline:10201543
23. Johanson NA, Liang MH, Daltroy L, et al. American Academy of Orthopaedic Surgeons lower limb outcomes assessment instruments. Reliability, validity, and sensitivity to change. J Bone Joint Surg Am. 2004;86-A(5):902–9. Medline:15118030
24. Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris questionnaire. Phys Ther. 1996;76(4):359–65, discussion 366–8. Medline:8606899

25. Alcock GK, Stratford PW. Validation of the lower extremity functional scale on athletic subjects with ankle sprains. Physiother Can. 2002;54(4):233–40.
26. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15. http://dx.doi.org/10.1016/0197-2456(89)90005-6. Medline:2691207
27. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50(8):869–79. http://dx.doi.org/10.1016/S0895-4356(97)00097-8. Medline:9291871
28. Stratford PW, Binkley JM, Riddle DL, et al. Sensitivity to change of the roland-morris back pain questionnaire: Part 1. Phys Ther. 1998;78(11):1186–96. Medline:9806623
29. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River (NJ): Prentice Hall Health; 2000.
30. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther. 1997;77(7):745–50. Medline:9225846
31. de Vet HCW, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033–9. http://dx.doi.org/10.1016/j.jclinepi.2005.10.015. Medline:16980142
32. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol. 2003;22(1):85–93. http://dx.doi.org/10.1002/uog.122. Medline:12858311
33. Pedhauzer EJ, Schmelkin LP. Measurement, design, and analysis: an integrated approach. Hillsdale (NJ): Lawrence Erlbaum; 1991.
34. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Out [Internet]. 2005 April [cited 2011 Aug 16];3:24. http://dx.doi.org/10.1186/1477-7525-3-23
35. Stratford PW, Spadoni G, Kennedy D, et al. Seven points to consider when investigating a measure's ability to detect change. Physiother Can. 2002;54(1):16–24.
36. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27(3 Suppl):S178–89. http://dx.doi.org/10.1097/00005650-198903001-00015. Medline:2646488
37. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care. 1990;28(7):632–42. http://dx.doi.org/10.1097/00005650-199007000-00008. Medline:2366602
38. Dennis RJ, Finch CF, Elliott BC, et al. The reliability of musculoskeletal screening tests used in cricket. Phys Ther Sport. 2008;9(1):25–33. http://dx.doi.org/10.1016/j.ptsp.2007.09.004. Medline:19083701
39. Finch E, Brooks D, Stratford PW, et al. Physical rehabilitation outcome measures: a guide to enhanced clinical decision making. Toronto: Canadian Physiotherapy Association; 2002.