Reliability and smallest real difference of the ankle lunge test post ankle fracture

David Simondsona,*, Kim Brocka, Susan Cottonb

a Physiotherapy Department, St Vincent's Hospital, PO Box 2900, Fitzroy 3065, Victoria, Australia
b Orygen Youth Health Research Centre, University of Melbourne, Carlton, Australia

Abstract

This study aimed to determine the reliability and the smallest real difference of the Ankle Lunge test in an ankle fracture patient population.

In the post immobilisation stage of ankle fracture, ankle dorsiflexion is an important measure of progress and outcome. The Ankle Lunge test measures weight bearing dorsiflexion, resulting in negative scores (knee to wall distance) and positive scores (toe to wall distance), for which the latter has proven reliability in normal subjects only.

A consecutive sample of ankle fracture patients with permission to commence weight bearing, were recruited to the study. Three measurements of the Ankle Lunge Test were performed each by two raters, one senior and one junior physiotherapist. These occurred prior to therapy sessions in the second week after plaster removal. A standardised testing station was utilised and allowed for both knee to wall distance and toe to wall distance measurement.

Data was collected from 10 individuals with ankle fracture, with an average age of 36 years (SD 14.8). Seventy seven percent of observations were negative. Intra and inter-rater reliability yielded intra class correlations at or above 0.97, \( p < .001 \). There was a significant systematic bias towards improved scores during repeated measurement for one rater (\( p = .01 \)). The smallest real difference was calculated as 13.8 mm.

The Ankle Lunge test is a practical and reliable tool for measuring weightbearing dorsiflexion post ankle fracture.

1. Introduction

The incidence of ankle fractures is higher than any other major upper or lower limb joint fracture in males aged 15–49 and either highest or second to wrist fractures only in women aged 15–69 (Singer et al., 1998). Range of ankle dorsiflexion is an important outcome measure in this condition and has been used consistently in studies of ankle fracture rehabilitation (Cimino et al., 1991; Hedstrom et al., 1994; van Laarhoven et al., 1996; Nightingale et al., 2006). Hancock et al. (2005) found weight bearing dorsiflexion, performed on ankle fractures on the day of cast removal, to be a consistent predictor of disability, perceived improvement and dorsiflexion range at six weeks and six months after cast removal. Failure to recover normal weight bearing dorsiflexion range has been observed in 53% of women and 32% of men two years after operatively treated ankle fracture dislocations (Lindsjo, 1985).

Given the high prevalence of ankle fracture, reported poor longer term outcomes and the predictive value of dorsiflexion measures, a clinically practical and reliable method to measure dorsiflexion post ankle fracture is required. Various techniques have been utilised including several instrumented techniques (Rowley et al., 1986; Moseley and Adams, 1991; Moseley et al., 2001), however these relatively complex systems may not be appropriate for clinical settings. Simple goniometry of the ankle in non-weight bearing has been shown to have low inter-rater reliability (Youdas et al., 1993). Measurements were found to be more reproducible using goniometry and weight bearing dorsiflexion rather than non-weight bearing in the normal population (Lindsjo et al., 1985).

The Ankle Lunge Test (ALT) is a simple linear measure requiring patients to perform weight bearing dorsiflexion by lunging their knee forward and beyond their toes (see Fig. 1). It uses inexpensive, easily obtainable equipment, does not require marking of bony landmarks, avoids error associated with goniometry and is in common use. Bennell et al. (1998), demonstrated very good
reliability for the ALT in a normal population, with a 95% confidence interval for between rater comparisons of 8 mm. They concluded that further research testing this measure is required in clinical populations. We were unable to identify any studies assessing reliability of the ALT or quantification of measurement error in clinical populations. Also, in clinical practice, many patients are unable to dorsiflex sufficiently to lunge their knee beyond the tip of their toe. Hancock et al. (2005) used knee to wall distance measurement (see Fig. 2) to enable a continuum of ALT range. This component of the ALT has not been tested for reliability, even though it has been utilised in published research on ankle fracture (Hancock et al., 2005; Lin et al., 2009).

The aim of this study was therefore twofold: to assess the consistency of the measurement process by evaluating inter-rater and intra-rater reliability of the ALT in ankle fracture, utilising both toe to wall and knee to wall measures, and to quantify the measurement error associated with the test by determining the smallest real difference, ie. the amount of change necessary to differentiate the effect of any intervention from measurement error. We hypothesised that the ALT would have good to very good intra-rater and inter-rater reliability, but would have a larger value for smallest real difference than the confidence intervals in a normal population.

1. Method

The study design was a within subject repeated measures design. Approval for this study was obtained from the Human Research and Ethics Committee at St. Vincent’s Hospital, Melbourne.

1.1. Participants

Patients were recruited consecutively at the fracture clinic by the duty physiotherapist. Inclusion criteria were referral to physiotherapy with a primary diagnosis of ankle fracture, permission to commence weight bearing as tolerated following plaster or similar immobilisation removal and a preference to attend rehabilitation at the St Vincent’s physiotherapy department. Patients were excluded from the study if they had significant other injury or co morbidities that would affect the testing procedure, or limited English below that required to read and complete the consent form. All patients who fitted the inclusion criteria were invited to join the study. Informed consent was received from the participants and their rights were protected.

Two physiotherapists participated in the study; a senior musculoskeletal therapist with three years musculoskeletal experience and a rotating junior therapist with four months musculoskeletal experience. A brief training session prior to the commencement of data collection allowed for familiarisation with the measure.

1.1.1. Measures

The ALT was performed at a designated testing station. The technique involved the use of a ruler set into the floor in front of a wall for toe to wall measures and a 15 cm set-square for knee to wall measures (see Figs. 1 and 2). The patient placed the foot of their injured ankle along the ruler, with their big toe against the wall and both their toe and heel on the centre line of the ruler. The patient was instructed to lunge their knee as close toward the wall as either stiffness or pain would allow. No control for pronation was provided. The patient could hold the wall for support if needed. The
therapist firmly held the patient’s heel down. At the point beyond which the patient’s heel was felt to be lifting, indicating the limit of range due to stiffness, or where the patient couldn’t tolerate further movement, indicating the limit of range due to pain, measurement was made. For patients who couldn’t reach the wall with their knee, the set-square, held against the wall by the therapist’s free hand, was used to read the distance from the anterior patella nearest to the wall (Fig. 1). This was recorded as a negative measure to the nearest millimetre. Patients who could reach the wall with their knee were asked to progressively move their toe further back from the wall on the ruler, repeating the lunge movement until the maximum distance at which they could tolerably lunge their knee to the wall without heel lift was found (Fig. 2). Measurement was made using the ruler from the tip of their toe to the wall. This was recorded as a positive measure to the nearest millimetre. The use of positive and negative measures avoids confusion in patient records where “toe to wall” can be interpreted as either the measure or the starting position of testing.

1.1.3. Procedures
Testing was performed in the second week following the commencement of weight bearing mobilisation and prior to any therapy provided on the day of attendance. A prior pilot study determined that the lunge measure could improve readily with therapy provided on the day of attendance. A prior pilot study was made. For patients who couldn’t tolerate further range due to stiffness, or where the patient couldn’t tolerate lunge their knee to the wall without heel lift was found (Fig. 2). Measurement was made using the ruler from the tip of their toe to the wall. This was recorded as a positive measure to the nearest millimetre. The use of positive and negative measures avoids confusion in patient records where “toe to wall” can be interpreted as either the measure or the starting position of testing.

A coin toss determined which therapist tested first for each patient. After each test, the patient was asked to step back away from the ruler, and a 10 s interval elapsed before the next was performed. This allowed adequate time for recovery from any symptoms consistently for all participants. A 10 min rest period was scheduled between the three ratings performed by the first therapist and those performed by the second therapist. The two therapists were blind to the each other’s ratings but not their own first and second measures on any given occasion.

1.1.4. Data analysis
Intra-rater and inter-rater reliability were assessed using intra-class correlations. For intra-rater reliability, ICC model (3,1) was used, assessing multiple scores from the same rater based on repeated measures ANOVA (Portney and Watkins, 2000). For inter-rater reliability, ICC model (2,2) was used, also based on repeated measures ANOVA, this time with rater as the independent variable (Portney and Watkins, 2000). Systematic bias was investigated using ANOVA and t tests. The smallest real difference (SRD) was evaluated using the standard error of measurement (SEM), derived as the square root of the error variance term of the ANOVA, and 95% confidence intervals (Stratford and Goldsmith, 1997; Stevenson, 2001), with the SRD being the product of the SEM, the tabled Z score for the 95% confidence interval (1.96) and the square root of 2. Smallest real difference scores were calculated utilising intra-rater assessments (SRD when assessments are conducted by the same rater) and inter-rater assessments (SRD when using more than one rater). SPSS version 15 was used for the statistical analysis.

2. Results
Ten patients, with average age 36 years (SD 14.8), six male, were recruited. Seven fractures could be defined according to Weber classification as B and one as C (Hughes et al., 1979), including three with associated tibial fractures. Of the remaining two, one ankle had bi-malleolar avulsions with an undisplaced talar fracture and the other an isolated medial malleolar fracture. Six ankles were managed by open reduction and internal fixation (ORIF). Demographic and medical details of the participants are provided in Table 1. Mean time since plaster removal was 1.5 weeks (SD 1). Rater 1 measured first on six out of the 10 occasions.

Table 2 shows the mean and standard deviation of scores for the ALT, 46 of 60 scores (77%) were negative. Figs. 3 and 4 show the distribution of scores for each rater. Intra-rater reliability (Table 2) yielded ICC’s of 0.99 for both therapists. Systematic bias was significant for Rater 1 (Table 2). Having considered this evidence of systematic bias, further analysis establishing the inter-rater reliability utilised the second measure taken, representing a clinically pragmatic choice of taking into account practice effects, but minimising the clinical time required for testing. Inter-rater reliability based on the second measure yielded an ICC of 0.97, (p < .001), with no evidence of systematic error (t(9) = 0.09, p = .93), see Fig. 5. The SRD was calculated for both intra-rater and inter-rater assessments. Same rater SRD was 8.3 mm for Rater One and 7.3 mm for Rater 2. Between rater SRD, based on the second measure, was 13.8 mm.

3. Discussion
This study indicates that application of the ALT at an early stage post plaster removal following ankle fracture, inclusive of a high proportion of negative measures, is reliable, irrespective of the experience of the clinician. This is the first study to provide reliability data for negative ALT measures, utilising a clinical sample with marked reductions in range of ankle dorsiflexion. It should be noted though, that the sample size in this study is small (n = 10) and therefore the results must be interpreted with caution.

This study demonstrated evidence of systematic bias, showing significant differences in repeated ratings by one rater, with scores improving up to a mean of 4.2 mm with three consecutive measures. It is likely that the repeated attempts at taking up the lunge position facilitated an increase in dorsiflexion range. Therefore it is important when using this measure to standardise the number of attempts the person will make on each testing occasion. In this study, due to this systematic bias, we chose to calculate
inter-rater reliability and SRD on the second of the three measures. This choice was made on a pragmatic basis, allowing the patient to have one practice attempt prior to measurement. It was considered that other alternatives, such as utilising the third measure, or the average of three measures would be too time consuming and onerous in the clinical scenario.

The SRD for multiple raters of 13.8 mm is higher than the inter-rater 95% confidence intervals for normal ankles (8 mm, Bennell et al., 1998) as hypothesised. As recovery takes place, reliability might be expected to approach values for normals as measurements become increasingly positive. The difference in degree of measurement error between the two studies may also relate to the observed improvement with consecutive testing in the clinical sample. The SRD derived from this clinical population should be applied to future ankle fracture research ahead of the normal ankle values.

The main limitation of this study is the lower than planned number of participants assessed by both assessors. The rate of recruitment was slower than anticipated and one rater resigned before the planned numbers were reached. This study’s sample included a wide range of fracture severity and management by both surgical and non-surgical methods, at an early stage post fracture removal. This may limit generalisability to latter stages of ankle rehabilitation. Although the sample is small, we consider that the participants assessed in this study demonstrated an adequate range of age, gender, fracture severity and management, representative of those that typically require significant physiotherapy intervention to complete their recovery. Additionally, testing was performed prior to therapy sessions, without any prior structured activity or warm up, which matches most closely what occurs clinically. This study compares measurements of an experienced clinician with an inexperienced clinician. This is also representative of typical outpatient physiotherapy staffing in a busy public hospital and avoids the risk of underestimating the value representing real change by using two experienced therapists.

It is important to note that the ALT is largely but not solely a measure of loaded talo-crural dorsiflexion. It is affected by available sub-talar eversion and midfoot pronation. No control for pronation was provided in this study, which is consistent with method in the normal study (Bennell et al., 1998) and in the predictive study of fractured ankles (Hancock et al., 2005). Furthermore it may be influenced by swelling, pain and weight bearing tolerance as well as soft tissue and joint compliance, thus making it a more global measurement of impairment (Bennell et al., 1998; Hancock et al., 2005) which increases its utility as an outcome to monitor effectiveness of treatment addressing these

### Table 2

<table>
<thead>
<tr>
<th>N</th>
<th>Mean (SD)</th>
<th>Stability of ratings</th>
<th>ICCb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALT in mm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 1</td>
<td>10</td>
<td>1. −30 (46.1) F(1.1,9.8) = 9.23, ICC (3,1) = 0.99,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. −27.6 (44) p = 0.01 p &lt; 0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. −24.3 (42.1)</td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>10</td>
<td>1. −28.3 (40) F(2,18) = 2.69, ICC (3,1) = 0.99,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. −27.8 (42.6) p = 0.095 p &lt; 0.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. −25.7 (44.3)</td>
<td></td>
</tr>
</tbody>
</table>

a Standard Deviation.
b Intraclass correlation.

Fig. 3. Distribution of scores for Rater One.

Fig. 4. Distribution of scores for Rater Two.

Fig. 5. Scores of Rater One and Rater 2 used in the inter-rater reliability analysis (second measure).
impairments, common to ankle fracture. Finally, being a linear rather than an angular measure does make it more sensitive at or near plantargrade, where 10 degrees of improvement would translate to a larger linear change than it would near full lunge. In our opinion the ALT, being both practical and reliable, is overall the most clinically useful method of dorsiflexion measurement in ankle fracture rehabilitation.

4. Key points

4.1. Findings

The findings of this study support the use of the ALT as a reliable measure, with a defined smallest real difference, best estimated as 13.8 mm in the ankle fracture population two weeks after plaster removal.

4.2. Implication

We recommend continued use of this test in clinical settings for assessment of ankle fracture and for evaluation of outcomes for individual patients. This simple test can continue to be utilised for research activities in clinical settings, potentially contributing to the development of the evidence base in the management of ankle fracture.

4.3. Caution

Small sample size and repeated measurement may affect the estimate of the SRD. This value may not be generalised to all stages of ankle fracture rehabilitation.

Conflict of interest

The authors received no financial support for this study. The authors have no conflict of interest.

Acknowledgements

Bridget Ward and Sarah O’Connell for their assistance.

Reference List


