The intra and inter-rater reliability of a modified weight-bearing lunge measure of ankle dorsiflexion

GRAFTON, Kate and O'SHEA, Simon

Published version

GRAFTON, Kate and O'SHEA, Simon (2013). The intra and inter-rater reliability of a modified weight-bearing lunge measure of ankle dorsiflexion. Manual Therapy, 18 (3), 264-268.

Title

The intra and inter-rater reliability of a modified weight-bearing lunge measure of ankle dorsiflexion.

Simon O'Shea MSc.

Kate Grafton MSc.

Email: simoncoshea@hotmail.com.

Correspondence: Physiotherapy Department, The Rotherham NHS Foundation Trust, Moorgate Road, Rotherham, South Yorkshire, S60 2UD. Tel 01709 424400

Keywords: Reliability, lunge, ankle, dorsiflexion.

Abstract

This study assessed the intra and inter-rater reliability of a modified weight-bearing lunge measure of ankle dorsiflexion range of movement. Thirteen healthy subjects were recruited. Each subject performed 3 repetitions of the lunging method with one rater and 3 more repetitions with a second rater within 30 minutes. The process was repeated within 3 hours. Intra-rater reliability results indicated excellent correlation of measurements (Intraclass Correlation Coefficients (ICCs) of 0.98 to 0.99). Standard Error of Measurement (SEM), 95% Limits of Agreement (LOA) and Coefficient of Repeatability (CR) calculations indicated suitably low ranges of measurement variance (SEM = 0.4cm, LOA = ±1.28 to ±1.47cm and CR = 1.21 to 1.35cm). Inter-rater reliability was also deemed excellent (ICC = 0.99, SEM = 0.3cm, LOA = ±0.83 to ±1.47cm, CR = 1.44cm). The modified lunge technique therefore demonstrates excellent intra and inter-rater reliability.

Text

Introduction

Suitable ankle dorsiflexion range of movement (DFR) is needed for efficient walking (Magee, 2008). Hypomobility of DFR is associated with pathologies including tendonopathies (Kaufman et al., 1999) and fractures (Agosta and Morarty, 1999); and restoring DFR is a common aim of rehabilitation following ankle fractures (Lin et al., 2009) and sprains (Collins et al., 2004). Consistent measurement before and after treatment is important so progress can be monitored.

Weight-bearing DFR measurements have demonstrated greater reliability and are more functionally orientated than non-weight bearing alternatives (Bennell et al., 1998; Aitkenhead 2002; Jones et al., 2005; Munteanu et al., 2009). Greater sensitivity (Bagget and Young, 1993) and superior cost and time effectiveness of functional weight-bearing methods have been claimed (Bennell et al., 1998; Jones et al., 2005).

A weight-bearing DFR measurement method that has demonstrated excellent reliability (Intraclass Correlation Coefficients (ICCs) of 0.97 to 0.99 for intra and inter-rater reliability respectively) involves lunging towards a wall (Bennell et al., 1998). The lunge is repeated up to 5 times to enable the foot to be moved away or towards the wall until the 'end range' is found.

An adapted version of the technique (Jones et al., 2005) involving pushing a moveable datum with the lunging knee has also shown good reliability (ICCs 0.82 to 0.99). However, use of customised equipment makes this technique less practical and more expensive.

A modified DFR measurement technique has been developed that can be viewed as a clinically simplified version of that proposed by Jones et al. (2005).

Instead of pushing a custom-made datum with the knee, the new technique uses the upright leg of a clinic table (see figure 1, table length 61cm, width 30cm and height 71cm). Three repetitions of the test are performed and the mean figure used. The benefits of this method above others are the speed of the test and simplicity of explanations to patients. Also, varied foot positioning may change the amount of pronation and subsequently affect DFR (Pope et al., 1998). With the modified technique the foot position can remain unaltered which improves standardisation of the technique. The modified technique may therefore be less prone to variation. Establishing the intra and inter-rater reliability is needed before this modified lunge DFR measurement technique can be recommended. Direct comparisons with Bennell et al. (1998) and Jones et al. (2005) would need specific equipment and more repetitions that may lead to mobilisation effects or prolong the study duration and introduce potential variance of DFR if measured on different days. The proposed lunge measure will therefore be compared to previous results of the aforementioned studies instead.

Method

Pilot Study

A pilot study (*n* = 5) was undertaken to refine instructions and inform a power calculation (Walter et al., 1998). The pilot study generated ICC scores of > 0.9. Type I and II error probability selected was 0.05 and 0.2 respectively. The ICC parameter was therefore set at 0.9 (Walter et al., 1998, table 2) giving a calculated sample size of thirteen.

Subjects

Thirteen volunteers (6 males, 7 females), mean age of 39 (standard deviation (SD) 14.5) and height of 168cm (SD 10.1) were recruited from staff at the Chesterfield Royal Hospital. Exclusion criteria (expanded from Munteanu et al., 2009) included acute or chronic lower limb pathology in the past year, previous lower limb surgery, neurological or balance deficits or an inability to perform or sustain a lunge for any reason.

Recruitment included verbal and emailed presentations to staff members. Written consent was gained and data was anonymised then securely stored. Sheffield Hallam University Research Ethics Committee gave ethical approval.

Raters

Two raters were used for all measurements. Rater 1 had 5 years clinical Physiotherapy experience and devised the modified technique. Rater 2 had 15 years of experience and was provided with a 15 minute training session to ensure standardisation between the raters.

Procedure

The full procedure and rationale is detailed in figures 1 and 2. Subjects looked forwards at all times and the tape measure was covered to blind the subjects from their performance. Raters measured many subjects in succession and had no access to previous measurements to minimise recall of data.

Data Analysis

Raw data was screened for anomalies. Bland and Altman plots (Bland and Altman, 1999), box plots and histograms assessed whether data was homoscedastic, normally distributed and not dependent upon the mean, which would affect statistical power (Atkinson and Neville, 1998; Bland, 2000).

Correlations were used to assess if age, height, or gender corresponded with measurements. Differences between the first and second measurement sessions, and between the two raters were evaluated using repeated ANOVA calculations. Post hoc statistical tests (Bonferroni) were performed where differences were identified.

Methods for assessing reliability have varied rationales and limitations; combinations of statistical methods are therefore suggested (Atkinson and Nevill, 1998; Rankin and Stokes, 1998).

ICC (3,k) was utilised (Shrout and Fleiss, 1979). Error range and repeatability was calculated with standard error of measurement (SEM), 95% confidence intervals (CI), 95% limits of agreement (LOA) and the coefficient of repeatability (CR) (British Standards Institute, 1979; Denegar and Bull, 1993; Atkinson and Nevill, 1998; Rankin and Stokes, 1998; Bland and Altman, 1999; Bland, 2000). 95% LOA demonstrate the range of measurement error within the sample and CR extrapolate a predictive figure for future measurement variance to 95% probability (British Standards Institution, 1979). The significance level was set at $p < 0.05$. SPSS version 16 software was used.

Results

Thirteen volunteers completed the study. No gender bias was evident. Histograms plus Bland and Altman plots confirmed that the data was homoscedastic (see figure 3). No dependence upon the mean and minimal measures beyond 95% LOA were evident (see figure 3) confirming a lack of anomalies or systematic bias.

Intra-rater Reliability

Excellent intra-rater correlation was found for rater 1 (ICC = 0.98) and rater 2 (0.99). See tables 1 and 2 for statistical analysis results. The level of error was also good (SEM = 0.4cm) for both raters. The spread of this error was small (95% CI = 0.8cm for both raters) and the maximum 95% LOA was ±1.47cm for rater 1 and ±1.28cm for rater 2, indicating a narrow band of difference between a raters first and second measurement session (see table 2 for all LOA data). CR indicated suitably small differences between repeated measurements (CR = 1.35cm for rater 1 and 1.21cm for rater 2).

Inter-rater reliability

Box plots demonstrated no significant anomalies (see figure 4) but rater 1 appeared to provide shorter measurements. Repeated ANOVA outcomes confirmed this and Bonferroni results show the second measurement session by rater 1 measured significantly lower (P = 0.01) than rater 2's second session with mean figures of 9.1cm (range 4.2 to 13.3) versus 9.5cm (range 4.8 to 14.1) respectively. No difference was demonstrated between rater 1 and rater 2 at the first session.

Despite the difference between the raters second session measurements, excellent inter-rater correlation was found (ICC = 0.99). The level of error was also good (SEM = 0.3cm). The spread of this error was small (95% CI = 0.6cm). LOA ranged from ±0.83 to ±1.47cm. A CR of 1.44cm also indicated a small difference between raters measurements.

Discussion

The results indicate the modified lunge DFR measurement technique is reliable with a healthy sample. ICC figures of 0.98 to 0.99 demonstrates the technique generates correlated repeated measurements (Bruton et al., 2000).

SEM and LOA figures evaluate the range of measurement variation and are recommended alongside ICCs (Denegar and Bull, 1993; Atkinson and Nevill, 1998; Rankin and Stokes, 1998; Bland and Altman, 1999). A maximum SEM of 0.4 (95% CI = 0.8) further supports the modified technique. The LOA indicate that a difference beyond ±1.47cm is needed to ensure that changes in measured distances are not the result of measurement variation with 95% confidence.

The CR provides the minimum detectable distance to 95% probability (British Standards Institute, 1979; Bland, 2000). CR figures of 1.21 to 1.35cm (intra-rater) and 1.44cm (inter-rater) are in accordance with LOA data. A difference of 0.03cm exists between the upper LOA and CR findings. If the conservative, larger figure is used, a measurement difference less than 1.47cm may be a result of measurement error. A difference greater than 1.47cm is deemed clinically significant and not attributable to measurement variability. Clinical responses to injury and treatments lead to changes that far exceed these ranges of ‘error’ and enable the technique to detect relevant changes. For example, a difference of 6.2cm has been noted between sprain injury patients and asymptomatic subjects (Collins et al., 2004).

Other studies used SEM (Bennell et al., 1998), CI or 75 percentiles (Hoch and McKeon, 2011) to provide ranges beyond which variance is thought to be absent. These methods do not provide 95% confidence or probability that the difference between two measurements is not attributable to error, unlike LOA (Rankin and Stokes, 1998; Bland and Altman, 1999) and CR (British Standards Institution, 1979). Previous lunge measurement reliability studies (Bennell et al., 1995; Jones et al., 2005) did not use a predictive statistic such as the CR but this enhances statistical analysis.

Claims of high reliability have been made by authors of other lunge DFR measurement techniques. Bennell et al. (1998) achieved similar ICCs (0.97 to 0.99) to the present study and a SEM (0.4 to 0.6) that was marginally larger, however no clear exclusion criteria was applied which may explain the greater variation of measures (SD = 3.7 to 4 compared with 2.8 in this study). Increased variation has been associated with inflated ICC figures (Denegar and Ball, 1993; Bland and Altman, 1999) because the calculation generates a relative index of variance. Increased variation in the range of measures enhances the power of such formulae to detect patterns in the calculations and vice versa (Mitchell, 1979; Haas, 1991; Atkinson and Nevill, 1998). Altering foot position every time may explain the greater variation.

The ICCs suggest the methods of Jones et al. (2005) are inferior to the modified technique (lowest figure of 0.66 compared to 0.98). Wider LOA of ±3.85cm compared to ±1.47cm with the modified technique strengthens this argument. Using specialist equipment also makes the datum method (Jones et al., 2005) more time intensive and expensive.

Better ICCs, SEM and LOA have been shown with the proposed modified DFR technique compared to alternatives (Bennell et al., 1995; Jones et al., 2005). The CR data gives further weight to these findings and a predictive confidence of 95%.

Limitations

Blinding of subjects to all results and the raters to previous measurements was undertaken but the potential for bias was high as the technique was devised by rater 1. The second rater, with no involvement with the technique or study, generated better ICC and LOA findings. This suggests the potential researcher bias was not present.

A significant inter-rater difference was found between the second session of measurements with rater 1 measuring shorter distances. This could be due to interpretations of heel lifting. Excessive grasping of the heel or over vigilance preventing pronation in an attempt to ensure strict standardisation may have altered the movement and explain a reduced score. One explanation for this may have been over-eagerness to limit any mobilisation effect as the second session was performed up to 3 hours later using subjects who were mobilising during this time. Kinematic and pressure sensor technology would enable assessment of this but would incur greater cost so was not available. Some studies have utilised an electromechanical lever (Aitkenhead, 2002) or a restraining strap placed over the mid foot region (Jones et al., 2005) but this was thought contrary to the clinically orientated aims of the present technique and difficult to standardise. Despite this discrepancy excellent inter-rater ICC results were evident.

Anecdotally, clinical use of the modified DFR measurement technique often results in patients being unable to touch the table leg with their knee due to hypomobility of the ankle. In these cases the shortest distance between the patella and the table leg is used as the measured distance. Similar methods have proven reliable with ankle fracture patients (Simondson et al., 2012). The use of the modified technique with patients requires further reliability assessment before it can be advocated widely.

Conclusion

This study demonstrated the proposed modified weight-bearing DFR lunge technique is reliable when used with a healthy sample. A difference greater than 1.47cm represents a meaningful difference beyond the variation of the technique.

References

Agosta J, Morarty R. Biomechanical analysis of athletes with stress fractures of the tarsal navicular bone: a pilot study. Australian Journal of Podiatric Medicine 1999; 33(1): 13-18.

Aitkenhead I. Ankle joint dorsiflexion assessment: the development of a new weight-bearing method. British Journal of Podiatry 2002; 5(2): 32-35.

Atkinson G, Nevill A. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine 1998; 26(4): 217-238.

Bagget B, Young G. Ankle joint dorsiflexion establishment of normal range. Journal of the American Podiatric Medical Association 1993; 85(5): 251-254.

Bennell K, Talbot R, Wajswelner H, Techovanich W, Kelly D. Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. Australian Journal of Physiotherapy 1998; 44(3): 175-180.

Bland M, Altman D. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999; 8(2): 135-160.

Bland M. An introduction to medical statistics, 3rd ed. Oxford University Press, 2000. p. 123-343.

British Standards Institute. Precision of test methods. Part 1: Guide for determination and reproducibility for a standard test method. British Standards Institute, London, 1979. p. 1-22.

Bruton A, Conway J, Holgate S. Reliability: What is it, and how is it measured? Physiotherapy 2000; 86(2): 94-99.

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. Manual Therapy 2004; 9(2): 77-82.

Denegar C, Ball D. Assessing reliability and precision of measurement: An introduction to intraclass correlation and standard error of measurement. Journal of Sport Rehabilitation 1993; 2(1): 35-42.

Haas M. Statistical methodology for reliability studies. Journal of Manipulative and Physiological Therapeutics 1991; 14(2): 119-132.

Hoch M, McKeon P. Normative range of weight-bearing lunge test performance asymmetry in healthy adults. Manual Therapy 2011; 16(5): 516-519.

Jones R, Carter J, Moore P, Wills A. A study to determine the reliability of an ankle dorsiflexion weight-bearing device. Physiotherapy 2005; 91(4): 242-249.

Kaufman K, Brodine S, Shaffer R, Johnson C, Cullison T. The effect of foot structure and range of motion on musculoskeletal overuse injuries. American Journal of Sports Medicine 1999; 27(5): 585-593.

Lin C, Moseley A, Herbert R, Refshauge K. Pain and dorsiflexion range of motion predict short and medium-term activity limitation in people receiving physiotherapy intervention after ankle fracture: An observational study. Australian Journal of Physiotherapy 2009; 55(1): 31-37.

Magee D. Orthopedic physical assessment, 5th ed. Saunders, Elsevier, 2008. p. 881.

Mitchell S. Interobserver agreement, reliability and generalizability in data collected in observational studies. Psychological Bulletin 1979; 86(2): 376-390.

Munteanu S, Strawhorn A, Landorf K, Bird A, Murley G. A weightbearing technique for the measurement of ankle dorsiflexion with the knee extended is reliable. Journal of Science and Medicine in Sport 2009; 12: 54-59.

Pope R, Herbert R, Kirwan J. Effects of ankle dorsiflexion range and pre-exercise calf muscle stretching on injury risk in army recruits. Australian Journal of Physiotherapy 1998; 44(3): 165-172.

Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clinical Rehabilitation 1998; 12: 187-199.

Shrout P, Fleiss J. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 1979; 86(2): 420-428.

Simondson D, Brock K, Cotton S. Reliability and smallest real difference of the ankle lunge test post ankle fracture. Manual Therapy 2012; 17(1): 34-38.

Walter S, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Statistics in Medicine 1998; 17(1): 101-110.

Table1: Means, SD and statistical results of each rater and both raters combined.

Rater	1st session Mean Distance & SD (cm)	2nd session Mean Distance & SD (cm)	ICC	SEM (cm)	95% CI (cm)	CR (cm)
1	9.1 (2.7)	9.1 (2.8)	0.98	0.4	-0.4 – 1.2	1.35
2	9.5 (2.9)	9.5 (2.9)	0.99	0.4	-0.4 – 1.2	1.21
1 + 2	9.3 (2.8)	9.3 (2.8)	0.99	0.3	-0.3 – 0.9	1.44

Table2: Differences between measurement sessions and between raters. R1S1 = rater 1, 1st measurement session ; R1S2 = rater 1, 2nd session; R2S1 = rater 2, 1st session; R2S2 = rater 2, 2nd session.

Rater & Session	Mean Difference	95% LOA	±
R1S1 Vs R1S2	-0.07	-1.57 to 1.43	1.47
R2S1 Vs R2S2	-0.04	-1.34 to 1.26	1.28
R1S1 Vs R2S1	0.48	-0.96 to 1.92	1.44
R1S1 Vs R2S2	0.45	-1.02 to 1.92	1.47
R1S2 Vs R2S1	0.49	-0.74 to 1.72	1.23
R1S2 Vs R2S2	0.45	-0.38 to 1.28	0.83

Figure 1a Starting position with foot placed on tape and toe against upright of table. (b) Final lunge position.

Figure 1b

Figure 2: Sequence of measurement sessions and rationale of standardised procedure.

Sequence	Rationale
The right ankle was used every time.	No differences noted between right and left healthy ankle DFR (Scharfbillig and Scutter 2004).
Subject's heel and great toe aligned directly over a strip of tape perpendicular to the table leg. Tip of great toe placed against the table leg.	To limit subtalar pronation and prevent compensatory methods that could alter DFR. Modified from Munteanu et al. (2009).
Subject permitted to place hands on a wall in front of the table.	Prevent loss of balance or apprehension and altered range during the lunge.
Subject asked to lunge forward and push the table as far away as possible without lifting the heel. To aim the knee over the 2nd toe.	To limit subtalar pronation and prevent compensatory methods that could alter DFR as per Munteanu et al. (2009) guidance.
Rater holds subjects heel to monitor any lift. A lift invalidates the result and is followed by another attempt.	Prevents findings representing knee/hip movements rather than ankle DFR. End range DFR achieved.
Distance from furthest point of foot i.e. great toe, to nearest point of the table leg measured to nearest 0.1cm with a tape measure.	Recognised, standardised and reflective of clinical use (Bennell et al., 1998; Jones et al., 2005; Munteanu et al., 2009; Vicenzino et al., 2001).
Subject blinded to measurement (tape measure covered by rater).	Prevents transient motivational effects (Hickey et al., 1992).
15 seconds recovery between repetitions.	Prevent the impact of fatigue. Too many lunges in quick succession without rest may increase flexibility.
Subject performs 3 repetitions with 1st rater. The mean figures were used for statistical analysis. A further 3 repetitions with 2nd rater within 30minutes then measured.	Repetitions in line with statistical guidance (Bland 2000, Hopkins 2009). 30 minute limit for 2nd assessment to prevent subjects range altering e.g. post exercise.
Process repeated within 1-3 hours. Each rater therefore providing two separate measurement sessions each.	1-3hrs thought to be long enough to prevent fatigue/learning/memory effects (Bennell et al., 1998) but brief enough to prevent impact of other variables e.g. injury.

Figure3: Bland and Altman plot demonstrating the difference between measures taken by rater 1 and rater 2 against actual measurements. Includes mean difference and 95% LOA.

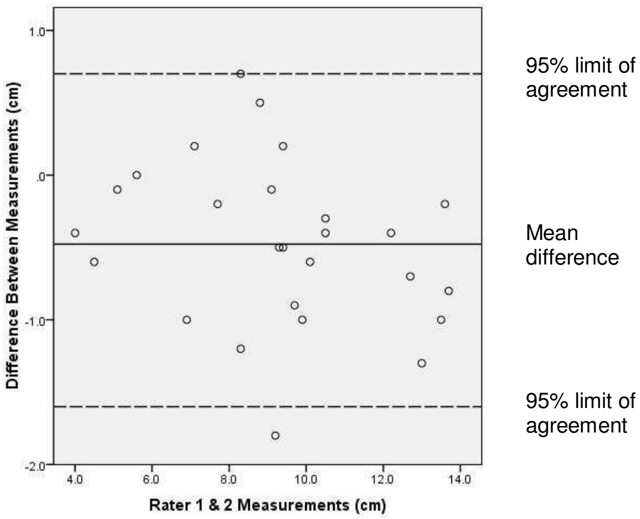

Figure 4: Box plot showing all measurements taken by rater 1 and rater 2. Means, upper/lower limits and interquartile ranges shown.