



Interrater, test-retest, and intersession reliability of a test designed to measure tibialis posterior strength with a hand-held dynamometer

Bergdal L^a, Svedmark Å^a, Lisbeth Brax-Olofsson^b, Tengman E^{a,*}

^a Dept. of Community Medicine and Rehabilitation, Section for Physiotherapy, Umeå University, Sweden

^b Dept. of Diagnostics and Intervention Section Orthopedics, Umeå University, Sweden

ARTICLE INFO

Keywords:

Physiotherapy
Foot strength
Test method
Measurement properties

ABSTRACT

Background: The tibialis posterior muscle has an important role both in stabilizing the foot and in inversion, plantar flexion, and adduction of the foot. Impaired function can lead to tibialis posterior dysfunction. A clinical test that can objectively measure tibialis posterior strength is warranted.

Objectives: The aim of this study was to investigate the interrater, test-retest, and intersession reliability of a test designed to measure tibialis posterior strength with a hand-held dynamometer.

Design: Interrater, between-day test-retest and intersession reliability.

Setting: University laboratory.

Participants: The participants comprised 20 healthy individuals (mean age 28.8 years, n = 10 women) without foot problems.

Method: A test was designed to test tibialis posterior strength with a hand-held dynamometer (HHD). The test was performed on two occasions 5–15 days apart and was carried out by two raters. The intraclass correlation coefficient (ICC), 95 % confidence interval, standard error of measurement (SEM), and minimal detectable change were calculated.

Results: Interrater reliability was good on both occasions (ICC: 0.769, 0.794), test-retest reliability was moderate for both raters (ICC: 0.671, 0.672), and intersession reliability was excellent (ICC: 0.934–0.967). However, the confidence interval had a large variation (–0.027–0.986) and the SEM was relatively high (2.356–3.863 N).

Conclusions: This test seems to be reliable, but has some limitations. The results suggest that the current version of the test could be used to compare strength between feet, but that further development of the test is needed to achieve increased interrater and test-retest reliability.

1. Introduction

The prevalence of foot and ankle pain in the adult population is estimated at 25–30 % (Gill et al., 2016; Hill et al., 2008; Thomas et al., 2011), with higher prevalence among women (Thomas et al., 2011). Previous foot pain, choice of shoes, pain in other joints, high body mass index (BMI), aging, and comorbidity all increase the risk of foot pain (Gill et al., 2016; Hill et al., 2008; Pita-Fernandez et al., 2017; Thomas et al., 2011). Patients with foot pain report lower self-assessed quality of life (Hill et al., 2008; Pita-Fernandez et al., 2017). Pes planus foot posture and pronated feet function are also associated with foot pain (Menz et al., 2013). An important muscle for the foot is the tibialis posterior muscle, which supports and stabilizes the medial longitudinal

arch and contributes to a normal gait pattern. It is also used for supination of the hind foot and plantar flexion of the ankle, and is involved in adduction and inversion of the foot (Kohls-Gatzoulis et al., 2004; Yao et al., 2015).

Like other muscles, the tibialis posterior can cause problems such as tendinosis, tendinitis, ruptures, and problems due to overuse or excessive pronation of the foot. These issues are usually located in the tendon area around the medial malleolus, and can lead to posterior tibialis tendon dysfunction (PTTD). PTTD is primarily established as a diagnosis through clinical examination, but often remains undiagnosed (Kohls-Gatzoulis et al., 2004; Ross et al., 2017; Yao et al., 2015). Impaired function and non-treated tibialis posterior dysfunction can lead to a flat foot deformity (Yao et al., 2015), which is characterized by

* Corresponding author. Umeå University, Department of Community Medicine and Rehabilitation 90187 Umeå, Sweden.

E-mail addresses: lisabergdahl@hotmail.com (B. L.), asa.svedmark@umu.se (S. Å.), lisbeth.brax.olofsson@regionvasterbotten.se (L. Brax-Olofsson), eva.tengman@umu.se (T. E).

<https://doi.org/10.1016/j.jbmt.2025.10.054>

Received 30 January 2025; Received in revised form 26 October 2025; Accepted 30 October 2025

Available online 1 November 2025

1360-8592/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

medial rotation and plantar flexion of the talus, eversion of the calcaneus, collapsed medial arch, and abduction of the forefoot (Pita-Fernandez et al., 2017; Yao et al., 2015).

There is a lack of clear guidelines for clinical assessment and physiotherapeutic treatment of PTTD (Rhim et al., 2022; Ross et al., 2018a; Yao et al., 2015). Ross et al. (2018b) found that exercise intervention in studies is often combined with orthoses and/or stretching, and that the performance of the exercise in the studies is often poorly described; they concluded that there is a need for high-quality studies (Ross et al., 2018b). Previous studies have used exercises described as strengthening for the tibialis posterior muscle (Alvarez et al., 2006; Houck et al., 2015; Kulig et al., 2004, 2009a, 2009b). Despite this, there is no test designed to objectively measure tibialis posterior strength. Being able to clinically test and evaluate the strength of the tibialis posterior muscle would be valuable for both diagnostic and treatment purposes.

Isokinetic measurement is considered the gold standard for measuring muscle strength, but it also has disadvantages such as the cost of the equipment and difficulty of use in clinical practice (Stark et al., 2011). An alternative is a hand-held dynamometer (HHD) (Bohannon, 2006), which is a reliable and valid instrument for measuring muscle strength (Chamorro et al., 2017; Stark et al., 2011). Previous studies have tested foot strength via inversion, starting from a slightly plantar flexed foot position (Alfuth and Hahm, 2016; Burns et al., 2005; Spink et al., 2010). The difficulty when testing from that position is that more muscles are activated, such as the tibialis anterior. This was noted in a study by Houck et al. (2015), which used electromyographic (EMG) biofeedback to control the movement into producing more tibialis posterior activation (Houck et al., 2015). Previous studies have included a variety of descriptions of how the tibialis posterior strength is measured or tested, in terms of both the position of the foot and the position of the participant (Alam et al., 2019; Burns et al., 2005; Kulig et al., 2004; Yao et al., 2015). Further development of a test is needed.

A reliable measurement tool can support diagnosis and evaluation of clinical treatment for tibialis posterior issues. This would be useful for physiotherapists in clinical practice, as the posterior tibialis strength is currently often assessed using manual resistance or functional test such as weight-bearing toe raises. Manual resistance is difficult to monitor over time and tests such as the toe raise can be too demanding for many patients with tibialis posterior dysfunction. A test with a HHD may provide a more quantitative, specific and objective method for assessing tibialis posterior strength. To ensure the use of the test in clinical settings, it is essential to establish its reliability for determine whether the test yields consistent results across different examiners, repeated measurements, and sessions over time. The aim of the present study was therefore to investigate the interrater, test-retest, and intersession reliability of a test designed to measure tibialis posterior strength with a HHD.

2. Material and method

2.1. Study design

This was a methodological study investigating interrater, test-retest, and intersession reliability. The study was conducted in accordance with recommendations from the COSMIN Risk of Bias tool for studies investigating reliability (Mokkink et al., 2020). All participants were provided with written and oral information about the study, and gave their written informed consent according to the Declaration of Helsinki. The project was approved by the Swedish Ethical Review Authority (dnr 2023-07616-01).

2.2. Participants

Participants comprised a convenience sample of 20 individuals (10 women). Inclusion criteria were being aged between 18 and 50 years, and having no current foot problems. Exclusion criteria were being

diagnosed with rheumatoid arthritis, neurological disease, previous foot surgery, pain worse than 1 on a visual analogue scale (VAS) running from 0 to 10, major orthopaedic trauma, or other similar conditions that could affect the foot. Information about gender, age, height, weight, physical activity level according to the validated Frändin and Grimby activity scale (Grimby and Frändin, 2018), type of training, dominant foot, previous problems with the right foot, and rated pain according to a VAS were collected. The Foot Posture Index (FPI-6) was used for assessing standing foot position. The FPI-6 is clinical tool designed to classify foot posture across multiple anatomical planes, and consists of six assessment criteria: talar head position, supra- and inframalleolar curvature, calcaneus frontal plan position, talonavicular prominence, medial longitudinal arch congruence, and abduction/adduction of the forefoot on rear foot. Together the score classifies the foot-position in the categories pronated, supinated, or neutral foot position. FPI is known to be a reliable and valid measurement tool (Redmond et al., 2006).

2.3. Procedure

The data collection was conducted during February 2024. Tests were performed by two raters, neither of whom had previous experience of using a HHD, but both of whom had clinical experience as physiotherapists. One rater had four years of experience as a physiotherapist specializing in orthopaedics and sports medicine, and the other had about 10 years of experience as a physiotherapist. As part of designing the test protocol, both raters trained approximately 1 h in the use of the HHD, focusing both on practical application and specifically on testing the tibialis posterior. In addition, pilot testing was conducted to allow for targeted training in the test procedure. The test order was predetermined. The raters were blinded to each other's results, and the participants were not told their results.

The first test occasion started with the participant filling out the forms, and then rater 1 went through all the steps of the test protocol, performed the test, and noted the results. After this, the participant rested for about 5–10 min. Then the next rater assessed the participant's FPI, performed the second muscle test, and noted the results. The second test occasion was 5–15 days later, and started with the participant rating their pain (VAS 0–10) in order to ensure there had been no changes in pain since the previous occasion. We also asked whether the participants had felt anything in their foot in the days leading up to this second test occasion, and made a note of any symptoms that were mentioned. The raters then performed the tests in the same order as at the first occasion.

2.4. Apparatus and test protocol

The HHD used in the tests was a Hoggan microFET2 (Scientific LCC, Salt Lake City UT, USA). This apparatus is wireless, custom sized to fit in the palm of the hand, easy to handle, and designed to produce objective and reliable measurements of muscle strength. Two testing techniques are available “make” or “break”. The “make” technique were used, where the subject exerts force against the HHD while the tester holds still. The “make” technique has been reported to be more reliable as well as carrying a lower risk of injury. When testing, it is important for the examiner to have a consistently stable base while the patient pushes.

A test protocol was designed based on previous studies testing muscle strength with a HHD for the foot (Alfuth and Hahm, 2016; Burns et al., 2005; Katoh, 2022; Spink et al., 2010), and studies describing tibialis posterior activation in plantar flexion, inversion, and adduction (Houck et al., 2015; Kohls-Gatzoulis et al., 2004; Kulig et al., 2004, 2009b; Yao et al., 2015). EMG biofeedback was used during the development of the test protocol to validate the best position of the foot for activating the tibialis posterior with minimum activation of the tibialis anterior. As also described by Houck et al. (2015), electrodes were placed on the tibialis anterior muscle to provide visual and aural feedback while the participant performed the movement and to identify the position that resulted in minimal muscle activation of tibialis anterior.

Five pilot tests were performed. The test positions of the tester and participant were standardized and the HHD was placed against the first metatarsophalangeal joint see Fig. 1.

The raters followed the test protocol to give the same instructions to all participants. The rater informed the participant that the test should not be painful, that the muscle being tested was a relatively small muscle, and that when pushing against the HHD the force should come only from the foot; the intention was to reduce the risk of the participant pushing with their leg and/or upper body. The function and anatomical position of the tibialis posterior muscle were described, and the movement was passively guided by the rater. Each participant began with 2–3 practice attempts that did not use the maximum contractions, in which they were guided by the rater to achieve the correct movement. These practice attempts were followed by a 2 min break and then three attempts using maximum contractions. Each attempt was held for 3–5 s with about 30 s of rest between. The rater gave clear instructions for when the participant should start, by saying “press”, and continued to say “press, press, press” for the next 3–5 s, then finished by saying “stop”. No other feedback was given during the test. The rater asked how the participant’s foot felt during and after the test, and any symptoms were noted.

2.5. Data and statistical analysis

Version 28 of the Statistical Package for the Social Sciences (IBM SPSS Statistics, Armonk, New York, USA) was used for all statistical analyses except for SEM, and MDC where Microsoft excel were used. The Shapiro–Wilk test revealed that all the data were normally distributed. Mean and standard deviation (SD) were calculated. The mean of the two best attempts at the muscle strength test was used to analyse interrater and test-retest reliability, while for intersession reliability all three attempts were analyzed separately. Difference in measurement value (DIFF) was calculated to describe differences between raters on the same test occasion and differences within the same rater when testing 5–15 days apart. This was first measured at an individual level, where any negative values were converted to positive values for the purposes of analysis. Paired samples t-tests were used to compare strength values between raters and occasions. There were no missing data.

Relative reliability was examined in terms of the intraclass correlation coefficient (ICC) with 95 % confidence interval (CI). A two-way random effects model with absolute agreement and single measurement was used to analyse interrater reliability. A two-way mixed effects model with absolute agreement and single measurement was used to analyse test-retest reliability. For intersession reliability a two-way



Fig. 1. A) Position of the participant half sitting on a bench with arms supported, feet and ankle just outside the bench. B) Foot resting position in plantar flexion. C) The active and assisted movement of plantar flexion, inversion, and adduction of the foot for tibialis posterior activation. D) Starting position of the test, a few degrees from outer position from C and with the HHD placed against the first metatarsophalangeal joint. E) The rater in sitting position with the elbow at 90° and anchored to the side of the body. By holding the right elbow with the left arm, the arm was stabilized. F) The fixed position for both participant and rater before the test started.

mixed effects model with absolute agreement and multiple/average measurements was used (Koo and Li, 2016; Treveltham, 2017). ICC values range from 0.00 to 1.00, where 1.00 means strong reliability. General guidelines for calculating reliability define <0.50 as poor reliability, 0.50–0.75 as moderate reliability, 0.75–0.9 as good reliability, and ≥0.9 as excellent reliability (Koo and Li, 2016).

Standard error of measurement (SEM) was calculated as SEM = Standard Deviation * √1-ICC and interpreted in terms of the assessment of reliability within participants (Weir, 2005). The minimal detectable change (MDC) was calculated as MDC = 1.96 * SEM * √2 to provide an indication of the smallest difference that would reflect a true change (Weir, 2005).

3. Results

A total of 20 participants were included in the study, comprising 10 men and 10 women aged 20–41 years (mean: 28.8 years). No participants dropped out during the study. The participants' characteristics are presented in Table 1. Nine of the participants described previous problems with the right foot, such as ankle sprains, tendon pain, and plantar fasciitis. The previous injuries occurred 1–8 years ago. Some of the participants said that they felt some soreness after the test. Aside from two participants who rated VAS 1 on the first occasion, all participants rated VAS 0 on both occasions. During the tests, some participants experienced a cramping sensation in the tibialis posterior muscle, the sole of the foot, and/or the medial side of the foot. No other adverse event was reported.

3.1. Tibialis posterior strength

Tibialis posterior strength measurements are presented in Table 2. On both test occasions, rater 1 measured a higher mean value than rater 2. At test occasion 1, there was no significant difference between raters (p = 0.125), whereas at test occasion 2, Rater 2 recorded significantly lower strength values (p < 0.01). Both raters measured lower strength values on the second occasion than on the first (p = 0.030 and p < 0.01). Men showed greater strength than women.

3.2. Interrater reliability of tibialis posterior strength

Interrater reliability was good at both test occasions (ICC: 0.769–0.794). The 95 % CIs showed a greater variation at the second test occasion (0.150–0.936). The absolute reliability (SEM) and difference in measurement values (DIFF) showed similar values on both occasions, while the MDC was lower on the second test occasion (Table 3).

Table 1
Participant characteristics.

	All	Women	Men
Participants n	20	10	10
Age years mean ± SD	28.8 ± 6.8	26.6 ± 5.7	31.0 ± 7.4
Height (m) mean (SD)	1.74 ± 0.07	1.69 ± 0.05	1.79 ± 0.06
Weight (kg) mean (SD)	71.6 ± 9.2	66.9 ± 8.0	76.3 ± 8.1
BMI (kg/m ²) mean (SD)	23.5 ± 1.7	23.4 ± 1.8	23.8 ± 1.7
Physical activity ^a median (range)	6 (2)	5 (2)	6 (2)
Dominant ^b Right n (%)	16 (80 %)	10 (100 %)	6 (60 %)
FPI ^c n			
Normal	12	7	5
Pronated	5	3	2
High pronated	2	0	2
Supinated	1	0	1
Days between tests mean ± SD	7.1 ± 2.2	7.8 ± 2.6	6.4 ± 1.4

^a According to Frändin Grimby activity scale.

^b The foot you would play soccer with.

^c Foot Posture Index -FPI right foot.

Table 2
Tibialis posterior strength* (N) measured with a hand-held dynamometer.

Strength (N)	Rater 1		Rater 2	
	T1	T2	T1	T2
All (n = 20) mean ± SD	46.3 ± 12.7	41.4 ± 12.2	43.5 ± 11.4	35.0 ± 10.6
(min to max)	(24.2–74.7)	(21.1–64.5)	(22.4–72.3)	(20.2–53.8)
Men (n = 10) mean ± SD	52.0 ± 11.1	43.2 ± 8.7	46.2 ± 11.3	38.1 ± 9.1
(min to max)	(35.3–74.7)	(30.9–55.2)	(34.7–72.3)	(22.0–51.5)
Women (n = 10) mean ± SD	40.6 ± 12.1	39.6 ± 15.2	40.7 ± 11.3	31.9 ± 11.6
(min to max)	(24.2–57.0)	(21.1–64.5)	(22.4–58.5)	(20.2–51.5)

Strength values was analyzed from mean of the two best attempts. N, Newton; T1, Test occasion one; T2, Test occasion two; SD, Standard deviation; Min to max, Minimum to maximum.

Table 3
Interrater and test-retest reliability of tibialis posterior strength^a (N) measured with a hand-held dynamometer.

Interrater and test-retest reliability				
	ICC (95 % CI)	SEM	MDC	DIFF
Interrater reliability between Rater 1 and Rater 2				
T1 (n = 20)	0.769 (0.507–0.901)	2.569	14.815	6.5 ± 5.3
T2 (n = 20)	0.794 (0.150–0.936)	2.356	6.530	6.9 ± 5.1
Test-retest reliability				
Rater 1 (n = 20)	0.671 (0.314–0.858)	3.863	10.709	8.6 ± 6.0
Rater 2 (n = 20)	0.672 (-0.027–0.894)	3.803	10.541	8.5 ± 5.9

^a Strength values was analyzed from mean of the two best attempts. N, Newton; T1, Test occasion one; T2, Test occasion two; ICC, Intraclass correlation coefficient (two-way random effects model, absolute agreement, single measurement); 95 % CI, 95 % Confidence interval; SEM, Standard error of measurement, presented in newton; MDC, Minimal detectable change, presented in newton; DIFF, Difference in measurement values between raters on the same occasion and for all participants, expressed as mean (SD) newton.

3.3. Test-retest reliability of tibialis posterior strength

Test-retest reliability was moderate for both raters (ICC: 0.671, 0.672). Rater 2 had a greater variation of the 95 % CIs, including a negative value (-0.027–0.894). The absolute reliability (SEM), MDC, and difference in measurement values (DIFF) were similar between raters (Table 3).

3.4. Intersession reliability of tibialis posterior strength

Intersession reliability was excellent for both raters and both test occasions (ICC: 0.928–0.967). The 95 % CIs had a small variation (0.848–0.986) for both raters and both test occasions (Table 4).

4. Discussion

The tibialis posterior muscle plays an important role in the functioning of the foot, and so there is a need to develop objective and clinically feasible tests for evaluating its strength. Previous studies have used an HHD to test foot and ankle strength in inversion (Alfuth and Hahm, 2016; Burns et al., 2005; Spink et al., 2010), while the test in the present study combines plantar flexion, inversion, and adduction. The present test showed good interrater reliability, moderate test-retest reliability for both raters, and excellent intersession reliability for both raters. However, the confidence interval had a large variation, and the SEM, MDC, and difference in measurement values were all relatively high, which suggests that the results should be interpreted with caution.

In the present study, the tibialis posterior strength test showed good

Table 4
Intersession reliability of tibialis posterior strength* (N) measured with hand-held dynamometer.

Intersession reliability			
	ICC (95 % CI)	SEM	MDC
Rater 1			
T1 (n = 20)	0.934 (0.861–0.972)	1.907	5.285
T2 (n = 20)	0.967 (0.932–0.986)	0.889	2.465
Rater 2			
T1 (n = 20)	0.928 (0.848–0.969)	1.932	5.355
T2 (n = 20)	0.963 (0.923–0.984)	0.881	2.442

All three attempts were analyzed separately. N, Newton; T1, Test occasion one; T2, Test occasion two; ICC, Intraclass correlation coefficient (two-way mixed effects model, absolute agreement, multiple/average measurements); 95 % CI, 95 % Confidence interval; SEM, Standard error of measurement, presented in newton; MDC, Minimal detectable change, presented in newton.

interrater reliability within the same day of testing and for both test occasions, with a higher agreement on the second occasion. This suggests a learning effect for both raters and participants. The 95 % CI had a greater variety, especially for the second test occasion. A previous study with a similar setup also showed a higher ICC value for interrater reliability on the second test occasion and a greater variety in the 95 % CI on the first test occasion when testing inversion strength in three different body positions (Alfuth and Hahm, 2016). On the other hand, Spink et al. (2010) showed good interrater reliability for inversion strength on both test occasions, and less variation of CI. Spink et al. tested the foot in a supine position, with the distal part of the ankle held still and a slight plantar flexion in the foot (Spink et al., 2010). The differences in ICC values could be because the three-dimensional movement of the foot performed in the present study is more difficult to repeat, but also the small sample size.

Test-retest reliability was moderate for both raters, and the 95 % CI showed a greater variety, even including a negative ICC value for rater 2. A negative value could be an incorrect estimate, and may occur if the sample size is small (Liljequist et al., 2019). A slightly better result for interrater reliability could be due to a learning effect for raters and participants when the test procedure is repeated with only a short time in between. Previous studies examining test-retest reliability for inversion strength with a short rest or a rest of a few hours have shown good to excellent reliability (Burns et al., 2005), while another study that used a 1-day interval showed moderate reliability (Alfuth and Hahm, 2016). This supports the assumption of a learning effect when testing with a short time between test occasions. Spink et al. (2010) investigated test-retest reliability with a 1-week interval, finding good to excellent reliability (ICC: 0.87, 0.9) for both raters and a smaller variation in the 95 % CI (0.8–0.94) compared to our study.

Intersession reliability was excellent for both raters, and the variation of 95 % CI was small, indicating a better result when testing within the same rater and test occasion. The SEM, MDC, and difference in measurement values all showed better values compared to interrater and test-retest reliability. Alfuth et al. (Alfuth and Hahm, 2016) previously reported moderate intrarater reliability for inversion in side-lying, supine, and sitting positions, with ICC values that were generally lower than in our study. However, other studies (Burns et al., 2005; Spink et al., 2010) have shown consistently higher ICC values when measuring inversion strength, regardless of the position in which the participants were tested, and regardless of whether the focus was on intrarater, interrater, or test-retest reliability (Alfuth and Hahm, 2016; Burns et al., 2005; Spink et al., 2010). Overall, better ICC values in previous studies could be due to the different starting position of the foot and difficulties using a three-dimensional movement when testing for interrater and test-retest reliability.

Regarding the ICC values in the present study, it is likely that our

findings are transferable to other physiotherapists with similar experience and in similar contexts, as the interrater reliability was good, while the transferability for test-retest reliability may be more uncertain. High values for SEM, 95 % CI, MDC could be considered uncertainty. The intersession reliability was excellent, and there was much less variation of the SEM, 95 % CI, MDC, and difference in measurement values. This suggests that the test method can be used to compare strength between the right and left foot.

The present study was carried out on healthy, training-active participants with normal to slightly pronated foot position. Reliability studies are recommended to be conducted on healthy individuals who remain stable over the test period (Mokkink et al., 2010, 2020). We controlled for adverse events by asking each participant how their foot felt during and after the test, and by asking them on the second test occasion whether they had experienced any symptoms in the foot during the days since the first occasion. No adverse events occurred during testing, but some of the participants felt a cramping sensation during the test.

Methodologically, the present study has both strengths and shortcomings. A strength was that the test procedure was carried out the same way on both test occasions. The raters were not blinded to the participants, but raters were blinded to the previous results and did not discuss the results with each other. Moreover, the participants were not told their results. Another strength was the use of EMG biofeedback when determining the test position during the developing the test protocol. Another study that measured inversion strength reported much higher values (Spink et al., 2010) than our study, which may indicate activation of other muscles. We noted a higher tibialis anterior muscle activation when the starting position of the foot involved less plantar flexion. This was also described by Houck et al. (2015). A limitation of the present study was the inclusion of only 20 participants and only two raters which affects the power and robustness of the reliability analyses. The sample size was based on an ICC = 0.8, a desired confidence interval width (CIW) of 0.3, and k = 2 measurements, where a sample size of 24 individuals was recommended (Wolak, 2012). To improve the transferability, more participants and a third rater would have made the study more robust. Koo and Li (2016) recommend at least 30 participants and at least three raters for studies investigating reliability (Koo and Li, 2016). Another limitation was the raters' limited experience using a HHD in this specific setup, meaning that the reliability may have been improved with more practice. It was, moreover, difficult to determine whether the raters correctly kept the HHD still or accidentally happened to press against it; this could be one reason why rater 1 consistently measured higher strength values compared to rater 2. This issue could be addressed through additional training sessions prior to testing. For future development of the test, the use of straps and external supports may enhance its reliability and precision. A further limitation was the difficulty of standardising the starting position and ensuring the combined movement of plantar flexion, inversion, and adduction. It is more difficult to test strength in outer range of motion, and a three-dimensional movement is more difficult to repeat. Some of the participants had more difficulty in engaging the correct muscle, especially during the inversion; a similar problem was seen in a previous study (Houck et al., 2015). We also noted that some participants had difficulty to controlling leg adduction, which required correction during the pilot-tests.

When analysing mean values, we used the best two results out of three attempts, while other studies have used a mean from three attempts (Alfuth and Hahm, 2016; Burns et al., 2005; Spink et al., 2010). The choice of absolute agreement when calculating the ICC can be seen as a strength. Trevethan (2017) recommends using absolute agreement and not consistency when calculating the ICC. The use of absolute agreement shows how identical the results are to each other, although it usually results in a lower ICC (Trevethan, 2017). For interrater reliability, we wanted to examine whether different raters assigned the same results to the same participant, as this is important for

transferability (Koo and Li, 2016).

5. Conclusion

Reliable assessment of tibialis posterior strength is crucial for clinical decision-making. Given the limitations of current manual and functional tests, a test using a HHD may offer a more objective useful clinical test. The test designed to measure three-dimensional strength demonstrated good interrater and test-retest reliability, and the excellent inter-session reliability suggests that the test can be used for comparing strength between feet within the same rater and test occasion. However, there was a large variation in the reliability, and the study had several limitations; the results should therefore be interpreted with caution. Further refinement of the test design is recommended to enhance its psychometric properties. For improvement of transferability, more participants and three raters are recommended. More studies that investigate objective, reliable, and clinically useful measurements of tibialis posterior strength are warranted.

CRedit authorship contribution statement

Bergdal L: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Svedmark Å:** Writing – review & editing, Validation, Methodology, Conceptualization. **L. Brax-Olofsson:** Writing – review & editing, Methodology, Conceptualization. **Tengman E:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Patient consent

All participants were provided with written and oral information about the study, and gave their written informed consent according to the Declaration of Helsinki. The project was approved by the Swedish Ethical Review Authority (dnr 2023-07616-01).

Funding

This work was supported by funding from the Strategic Research Area, Health Care Science (SFO-V) Umeå University. The funders did not have any role in study design, etc.

Declaration of competing interest

The authors declare they have no competing interests.

References

- Alam, F., Raza, S., Moiz, J.A., Bhati, P., Anwer, S., Alghadir, A., 2019. Effects of selective strengthening of tibialis posterior and stretching of iliopsoas on navicular drop, dynamic balance, and lower limb muscle activity in pronated feet: a randomized clinical trial. *Physician Sportsmed.* 47, 301–311.
- Alfuth, M., Hahm, M.M., 2016. Reliability, comparability, and validity of foot inversion and eversion strength measurements using a hand-held dynamometer. *Int. J. Sports Phys. Ther.* 11, 72–84.
- Alvarez, R.G., Marini, A., Schmitt, C., Saltzman, C.L., 2006. Stage I and II posterior tibial tendon dysfunction treated by a structured nonoperative management protocol: an orthosis and exercise program. *Foot Ankle Int.* 27, 2–8.
- Bohannon, R.W., 2006. Hand-held dynamometry: adoption 1900–2005. *Percept. Mot. Skills* 103, 3–4.
- Burns, J., Redmond, A., Ouvrier, R., Crosbie, J., 2005. Quantification of muscle strength and imbalance in neurogenic pes cavus, compared to health controls, using hand-held dynamometry. *Foot Ankle Int.* 26, 540–544.
- Chamorro, C., Armijo-Olivo, S., De la Fuente, C., Fuentes, J., Javier Chiroso, L., 2017. Absolute reliability and concurrent validity of hand held dynamometry and isokinetic dynamometry in the hip, knee and ankle joint: systematic review and meta-analysis. *Open Med.* 12, 359–375.
- Gill, T.K., Menz, H.B., Landorf, K.B., Arnold, J.B., Taylor, A.W., Hill, C.L., 2016. Predictors of foot pain in the community: the North West Adelaide health study. *J. Foot Ankle Res.* 9, 23.
- Grimby, G., Frandin, K., 2018. On the use of a six-level scale for physical activity. *Scand. J. Med. Sci. Sports* 28, 819–825.
- Hill, C.L., Gill, T.K., Menz, H.B., Taylor, A.W., 2008. Prevalence and correlates of foot pain in a population-based study: the North West Adelaide health study. *J. Foot Ankle Res.* 1, 2.
- Houck, J., Neville, C., Tome, J., Flemister, A., 2015. Randomized controlled trial comparing orthosis augmented by either stretching or stretching and strengthening for stage II Tibialis posterior tendon dysfunction. *Foot Ankle Int.* 36, 1006–1016.
- Katoh, M., 2022. Test-retest reliability of isometric ankle plantar flexion strength measurement performed by a hand-held dynamometer considering fixation: examination of healthy young participants. *J. Phys. Ther. Sci.* 34, 463–466.
- Kohls-Gatzoulis, J., Angel, J.C., Singh, D., Haddad, F., Livingstone, J., Berry, G., 2004. Tibialis posterior dysfunction: a common and treatable cause of adult acquired flatfoot. *BMJ* 329, 1328–1333.
- Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163.
- Kulig, K., Burnfield, J.M., Requejo, S.M., Sperry, M., Terk, M., 2004. Selective activation of tibialis posterior: evaluation by magnetic resonance imaging. *Med. Sci. Sports Exerc.* 36, 862–867.
- Kulig, K., Lederhaus, E.S., Reischl, S., Arya, S., Bashford, G., 2009a. Effect of eccentric exercise program for early tibialis posterior tendinopathy. *Foot Ankle Int.* 30, 877–885.
- Kulig, K., Reischl, S.F., Pomrantz, A.B., Burnfield, J.M., Mais-Requejo, S., Thordarson, D. B., Smith, R.W., 2009b. Nonsurgical management of posterior tibial tendon dysfunction with orthoses and resistive exercise: a randomized controlled trial. *Phys. Ther.* 89, 26–37.
- Liljequist, D., Elfving, B., Skavberg Roaldsen, K., 2019. Intraclass correlation - a discussion and demonstration of basic features. *PLoS One* 14, e0219854.
- Menz, H.B., Dufour, A.B., Riskowski, J.L., Hillstrom, H.J., Hannan, M.T., 2013. Association of planus foot posture and pronated foot function with foot pain: the Framingham foot study. *Arthritis Care Res.* 65, 1991–1999.
- Mokkink, L.B., Boers, M., van der Vleuten, C.P.M., Bouter, L.M., Alonso, J., Patrick, D.L., de Vet, H.C.W., Terwee, C.B., 2020. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med. Res. Methodol.* 20, 293.
- Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* 63, 737–745.
- Pita-Fernandez, S., Gonzalez-Martin, C., Alonso-Tajes, F., Seoane-Pillado, T., Pertega-Diaz, S., Perez-Garcia, S., Seijo-Bestilleiro, R., Balboa-Barreiro, V., 2017. Flat foot in a random population and its impact on quality of life and functionality. *J. Clin. Diagn. Res.* 11, LC22–LC27.
- Redmond, A.C., Crosbie, J., Ouvrier, R.A., 2006. Development and validation of a novel rating system for scoring standing foot posture: the foot posture index. *Clin. Biomech.* 21, 89–98.
- Rhim, H.C., Dhawan, R., Gureck, A.E., Lieberman, D.E., Nolan, D.C., Elshafey, R., Tenforde, A.S., 2022. Characteristics and future direction of tibialis posterior tendinopathy research: a scoping review. *Medicina (Kaunas)* 58.
- Ross, M.H., Smith, M., Plinsinga, M.L., Vicenzino, B., 2018a. Self-reported social and activity restrictions accompany local impairments in posterior tibial tendon dysfunction: a systematic review. *J. Foot Ankle Res.* 11, 49.
- Ross, M.H., Smith, M.D., Mellor, R., Vicenzino, B., 2018b. Exercise for posterior tibial tendon dysfunction: a systematic review of randomised clinical trials and clinical guidelines. *BMJ Open Sport Exerc. Med.* 4, e000430.
- Ross, M.H., Smith, M.D., Vicenzino, B., 2017. Reported selection criteria for adult acquired flatfoot deformity and posterior tibial tendon dysfunction: are they one and the same? A systematic review. *PLoS One* 12, e0187201.
- Spink, M.J., Fotoohabadi, M.R., Menz, H.B., 2010. Foot and ankle strength assessment using hand-held dynamometry: reliability and age-related differences. *Gerontology* 56, 525–532.
- Stark, T., Walker, B., Phillips, J.K., Fejer, R., Beck, R., 2011. Hand-held dynamometry correlation with the gold standard isokinetic dynamometry: a systematic review. *Pm R* 3, 472–479.
- Thomas, M.J., Roddy, E., Zhang, W., Menz, H.B., Hannan, M.T., Peat, G.M., 2011. The population prevalence of foot and ankle pain in middle and old age: a systematic review. *Pain* 152, 2870–2880.
- Trevelyan, R., 2017. Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Serv. Outcome Res. Methodol.* 17 (2), 127–143.
- Weir, J.P., 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Condit Res.* 19, 231–240.
- Wolak, M.E., 2012. Guidelines for estimating repeatability. *Methods Ecol. Evol.* 3, 129–137.
- Yao, K., Yang, T.X., Yew, W.P., 2015. Posterior Tibialis Tendon dysfunction: overview of evaluation and management. *Orthopedics* 38, 385–391.