ORIGINAL ARTICLE

Reliability of a device measuring triceps surae muscle fatigability

M Haber, E Golan, L Azoulay, S R Kahn, I Shrier

Br J Sports Med 2004;38:163–167. doi: 10.1136/bjism.2002.002899

See end of article for authors' affiliations

Correspondence to:
Dr Ian Shrier, LadyCentre
for Clinical Epidemiology
and Community Studies,
Davis Institute for Medical
Research, SMBD-Jewish
General Hospital, 3755
Cote Sainte-Catherine
Road, Montreal, Quebec,
Canada H3T 1E2;
ian.shrier@mcgill.ca

Accepted 26 March 2003

Objective: To examine the test–retest reliability of a protocol using an apparatus designed to standardise the standing heel rise test for the triceps surae muscle.

Subjects: 40 healthy subjects volunteered to test short and medium term test–retest reliability (group SM, median age 24 years), and a convenience sample of 38 subjects with a history of unilateral deep vein thrombosis (DVT) volunteered to test long term test–retest reliability (group L, median age 52 years).

Design: Subjects carried out 23 heel rises per minute until either the pace or the height could no longer be maintained. Group SM subjects repeated the test 30 minutes later (short term), and again 48 hours later (medium term). Subjects in group L did the test on the unaffected leg, and repeated the test one week later (long term).

Results: The median number of heel rises achieved per trial in group SM was 34 (range 16 to 120). The intraclass coefficient (ICC) was 0.93 (SEM 2.1) for both 30 minute and 48 hour test–retest reliability. In group L, the median number of heel rises was 27 (range 9 to 97), with ICC 0.88 and SEM 3.4.

Conclusions: The apparatus is a simple and inexpensive standardised tool that reliably measures triceps surae fatigability in subjects with no current injury. Future research should assess its use in injured patients.

Although muscle fatigue and weakness (defined as the inability to generate maximum force when carrying out repeated contractions, which is reversed with rest) is believed to be one of the main risk factors for injury during sports,^{1–3} there are few clinical tests for muscle fatigability. This is especially important in rehabilitation after injuries, where muscle strength and endurance are often affected at the time of the injury or during the rehabilitation period.

Laboratory tests such as power spectrum analysis,^{4,5} twitch interpolation,⁶ and isokinetic testing⁷ can all measure muscle fatigue precisely but are not readily available to the clinician because they are too costly or time consuming, or because they involve invasive equipment. A simple and practical method for measuring muscle fatigability would be an important clinical tool.

To assess fatigability of the triceps surae muscle, some investigators have used a standing heel rise test,⁸ an inexpensive and non-invasive indicator of calf muscle fatigability. However the position of the lower extremity, the height to be lifted, and the cadence of the exercise are all factors that affect the workload imposed and would therefore affect the reliability of such a test. In addition, a satisfactory criterion for the termination of a heel rise test must be established.

In the present study, we considered fatigue to have occurred when a subject was no longer able to maintain a required workload.⁹ More recently, this definition has been given the term “task failure.”¹⁰

The objective of this study was to assess the reliability of a protocol using an apparatus specifically designed to standardise the standing heel rise test for triceps surae muscle fatigability on a healthy group of subjects without a current injury.

METHODS

Subjects

We used a longitudinal test–retest design. To test short term (30 minute) and medium term (48 hour) test–retest reliability, we used a convenience sample of 40 healthy

subjects (group SM), the only exclusion criterion being a current injury that could cause pain during a heel rise exercise test. To test long term reliability (seven day), we used the unaffected leg of a convenience sample of 38 patients with a diagnosis of unilateral deep vein thrombosis (DVT) established at least one year before at our hospital (group L). These patients were recruited from a parallel study that was designed to evaluate exercise induced symptoms in this population,¹¹ and as part of that study, we were already measuring flexibility and treadmill endurance on tests done one week apart. We simply added this test of calf endurance, and it was done before the treadmill test, so that fatigue would not be a confounding factor. The exclusion criteria for the parallel exercise study were bilateral DVT (which would preclude the use of the unaffected leg as a control), symptomatic pulmonary embolism (potentially limiting the exercise capacity independent of leg symptoms), and refusal to provide informed consent.

Tool description

The apparatus (fig 1) consists of a rod and a foot positioning device attached to a platform, both of which are adjustable. Before each trial, the apparatus is first adjusted to limit the height of each heel rise to 5 cm as follows: first, the subject stands barefooted on one leg with the knee fully extended; second, the subject raises the heel until the navicular bone touches the rod. The foot positioning device or the rod or both are adjusted to limit the height of each lift to 5 cm (with the distal end of the subject's toes touching the foot positioning device at all times).

Fatigue test

We developed our protocol on a separate group of six healthy subjects with the objective of creating fatigue within two to three minutes, because we felt this was the right balance between measuring the appropriate type of endurance and the amount of time available for clinicians and large epidemiological studies.

Figure 1 The Haberometer. Before each trial the foot positioning device and rod are adjusted to limit each heel rise to the desired height. The subject stands on the platform barefooted with the knee fully extended and the contralateral leg suspended in air. Following a metronome, the subject does heel rises, touching the rod with the navicular bone each time until either the pace or the height can no longer be maintained, at which time the test is terminated. The pictures of the subject show her foot positioned with the distal end of her toes touching the foot positioning device (lateral view), and the navicular bone touching the rod during the maximum height of the heel lift (anterior view). The subject is allowed to place one hand lightly on a wall to improve balance.

While maintaining the knee of the testing leg fully extended throughout and the contralateral leg suspended in air, each subject raised their bare heel (no shoes or socks) until the navicular bone touched the rod, and then lowered it back to the platform at a rate of 46 beats/min (23 lifts/min). At the sound of the first beat of a metronome, the subject lifted the heel to the rod. This position was maintained until the sound of the second beat, at which time the subject lowered the heel and awaited the next beat. Throughout the entire trial each subject was permitted to place one hand flat against a wall for balance but not for support.

Subjects continued to raise and lower the heel until one of either two possible outcomes occurred:

- the pace could no longer be maintained;
- the height of 5 cm could no longer be attained.

Subjects who experienced soreness were encouraged to continue, but subjects who experienced pain were told to stop. A research assistant was present at all times to correct

any errors of technique and to record the number of repetitions.

To examine short term test-retest reliability (group SM), tests were repeated 30 minutes after the first test. To determine intermediate term test-retest reliability, group SM subjects returned two days afterwards to repeat the test two more times, following the same protocol. All of a given subject's trials were carried out at similar times of day to prevent the influence of circadian rhythms on muscular performance.¹² To examine long term test-retest reliability (group L), we followed the same protocol as the short term and medium term test-retest reliability, and tested the unaffected leg of group L on two separate occasions no less than a week apart.

Although three different research assistants conducted the tests on different individuals, the same research assistant tested each subject during multiple trials for that individual.

Statistics

Data analysis was done using Statview computerised software (SAS institute Inc, Cary, North Carolina, USA). As there is no single acceptable method of determining reliability, we report the intraclass correlation coefficient (ICC), standard error of measurement (SEM), coefficient of variation (CV), and limits of agreement.¹³

We calculated the ICC under the assumption of a two way random effects model, ICC(2,1).¹⁴ This method was chosen because the same examiner tested each subject, and different examiners never tested the same subject.¹⁴

The SEM was calculated according to Atkinson and Nevill¹⁵: $SEM = SD \sqrt{1 - ICC}$. The standard error of measurement is also known as the typical error,¹⁵ and represents the variation between a single measure (which includes some error) and the "true" score for the individual.¹⁵⁻¹⁷

We calculated the CV by averaging the results from individuals—that is, the average of [SD of two tests]/[mean of two tests] for each individual.¹³ The CV is the standard deviation expressed as a percentage. It is similar to the SEM but more appropriate when the error associated with a measurement increases as the value of the measure increases.

Another measure of reliability is the limits of agreement.¹⁸ Although gaining popularity because it represents the range of values to be expected 95% of the time if the test is repeated,¹³ others argue that it provides a biased result if there are fewer than 25 subjects (not relevant for the current study) or is simply 2.8 times the SEM when numbers are large.^{15, 17} In our population, the raw data showed signs of heteroscedasticity (that is, a positive correlation between absolute differences and individual means, suggesting that variation increased as the number of repetitions increased) which was reduced with log transformation. We therefore calculated limits of agreement for each pair of observations based on the log transformed data, and back transformed them to obtain the $\pm$ percentage error, which is more useful for the clinician.¹³

RESULTS

Of the 40 subjects in group SM (21 women, 19 men; median age 24 years, range 17 to 63), two could not return for the second day trials for personal reasons. Their data were used in the short term test-retest reliability analysis. Only one subject in the preliminary study complained of soreness that lasted a few days following testing.

The median number of lifts per trial among the subjects for short term and medium term test-retest reliability was 34 (range 16 to 120), and the test required approximately two minutes to complete in the majority of trials. Figure 2A is a plot of the results for each subject for the first two trials done 30 minutes apart on the first day. There was excellent

Figure 2 (A) Short term (30 minute) test-retest reliability: plot of the number of repetitions achieved on the second trial of the first day against the first trial of the same day, completed 30 minutes apart, by each subject (group SM, $n = 40$). The straight diagonal line represents the line of identity. The ICC is 0.93 (SEM 2.1), indicating excellent test-retest reliability. (B) The differences between the logs of the two first day trials are plotted against each subject's mean for the log of the two tests. The 95% limits of agreement are also presented on the plot. The point distribution suggests that there is no relation between the difference and the mean of the first day trials when plotted on a log-log scale. (C) Intermediate term (48 hour) test-retest reliability: plot of the number of repetitions performed on the first trial of the second day against the first trial of the first day, at least 48 hours later, by each subject (group SM). The ICC is 0.93 (SEM 2.1), indicating excellent test-retest reliability. Note that two subjects could not return for the second day trials for personal reasons ($n = 38$). (D) The differences between the log of the first trial of the second day and the log of the first trial of the first day are plotted against each subject's mean for the logs of the two tests. The 95% limits of agreement are also presented on the plot. The point distribution suggests that there is no relation between the difference and the mean of the trials when plotted on a log-log scale.

agreement (ICC = 0.93, SEM for the difference between the two first day trials = 2.1 repetitions). The mean difference between trials was 0.11 repetitions and the CV was 9%. The limits of agreement for the two different trials were $\pm 20\%$.

Figure 2C is plot of the results of each subject's first trial of day 2 against the first trial of day 1. The ICC of 0.93 and the SEM of 2.1 are the same as when two trials are undertaken 30 minutes apart. The mean difference between trials was -0.14 repetitions and the CV was 18%. The limits of agreement are shown in fig 2D and were $\pm 40\%$.

Within group SM, there were three individuals who completed more than 80 repetitions, whereas most individuals completed fewer than 60 repetitions. Although there is no apparent reason to remove these individuals, these outlying values may artificially increase the ICC. However, an analysis excluding these individuals still yielded an ICC of 0.85 (SEM = 2.3) for the 30 minute test-retest and 0.79 (SEM = 3.1) for 48 hour test-retest, both of which still suggest very high reliability.

Of the 38 subjects tested in group L (16 women, 21 men; median age 51 years, range 25 to 76), only one complained of pain and soreness that lasted a few days following the first test and declined a second trial ($n = 37$). Overall, the median number of lifts was 27 (range 7 to 97). Figure 3A is a plot of the results for the two trials completed no less than a week apart. There is excellent agreement (ICC = 0.88, SEM = 3.4).

The mean difference between trials was -0.63 repetitions and the CV was 15%. The limits of agreement are shown in fig 3B and were $\pm 29\%$.

DISCUSSION

According to the ICC results, the results of this study suggest that the apparatus has very good short term, medium term, and long term test-retest reliability.¹⁹ The SEM was approximately two repetitions for all trial comparisons, which can be interpreted as the typical variation between a single test measure and the "true" score for that individual. However, this is an underestimate because the limits of agreement analysis suggested that there was heteroscedasticity (as is true for many sports medicine measurements¹³), and therefore the variation is expected to be more than two repetitions for subjects who can perform a large number of repetitions. In this case the CV may be more appropriate, and in our experiment the CV suggests that the error would typically be expected to be between 9% and 18%. It also suggests that if the experiment were repeated many times, the repeated test would be within 9% and 18% of the original test 52% of the time.^{16,17} Finally, the limits of agreement suggest that if the experiment were repeated many times, the repeated test would be within 20% and 40% of the original test 95% of the time.^{16,17} The choice between these measures remains controversial at the present,^{13,15-17} but 95% limits are likely

Figure 3 Long term (minimum seven day) test-retest reliability (using the unaffected leg of patients with a previous history of deep vein thrombosis). (A) Plot of the number of repetitions for trial 2 against the number of repetitions for trial 1 (group L, $n = 37$). The straight diagonal line represents the line of identity. The ICC is 0.88 ($SEM = 3.4$), indicating excellent test-retest reliability. (B) The differences between the logs of the two trials with the unaffected leg are plotted against each patient's mean for the logs of the two tests. The 95% limits of agreement are also presented on the plot. The point distribution suggests that there is no relation between the difference and the mean of the trials when plotted on a log-log scale.

to be too strict unless a high degree of confidence in a single measure is extremely important (for example, drug testing).¹⁶

Because of its simplicity, the apparatus may be useful to measure calf fatigability in the clinic, and in research using large populations. The large range of repetitions recorded between subjects suggests that it is useful for healthy individuals with any level of calf function. In addition, of the 84 subjects (40 group SM, 38 group L, six controls used during pretesting), only two complained of soreness and this only lasted for a couple of days following the test. Moreover, one of these individuals participated in a strenuous hiking trip the day following the test and even then his pain only lasted a few days. Nevertheless, we now recommend avoiding any strenuous activity involving the triceps surae for the few days following the test.

Currently, the most common method of examining muscular performance is with commercially available isokinetic dynamometers.⁷ These devices generate large quantities of computed measurements that combine several variables relating to muscle force, power, and fatigability. Although most commercial dynamometer systems are intrinsically accurate in their measurements,⁷ the complexity of the equipment and of the movements involved require that both tester and subject receive thorough and time consuming instruction and familiarisation with the testing procedures before use in order to achieve high test-retest reliability.⁷ Also, they are extremely expensive and as a result are not readily available to the clinician or for use in large epidemiological studies. Although muscle fatigability can also be determined indirectly by other means such as electromyography and twitch interpolation techniques, their use may not always be convenient. In contrast, our device is inexpensive to build, and its simplicity allows it to be used in "field" conditions outside the laboratory. As there are no complex movements required, no learning effect was expected or observed.

Our measure of fatigue was task failure, which can be influenced by motivation. During our test, subjects were not told how many repetitions they achieved. Although it is possible that they counted these themselves, subjects found this quite difficult to do during our pilot testing. Further, even if they remembered for the 30 minute test-retest, we feel it is unlikely that they would have remembered for the 48 hour test-retest, and even more unlikely for the seven day test-retest (especially as these latter subjects underwent two

hours of testing for the parallel study after carrying out the calf fatigability measure). As the reliability for the 48 hour and seven day test-retest measures was also very good, we do not believe that recall was a significant factor for motivation in our study.

Soreness can also affect motivation. We asked subjects to stop once they reported pain, but encouraged them to continue if there was just soreness. As some individuals have difficulty understanding the difference, this could have affected our measures. However, the high reliability of the measure within individuals suggests that the magnitude of any associated error is small. Finally, we used different groups of subjects for the short/medium term reliability and the long term reliability parts of the study, suggesting that the results were not specific to one convenience sample population.

In a previous study incorporating a similar standing heel rise exercise,⁸ other researchers increased the amount of work with each lift by using an angled platform which caused increased dorsiflexion of the foot. However, this places the muscle in a stretched position and could increase the risk of injury. We therefore suggest increasing work by increasing the cadence or lift height, or by the use of weights. Furthermore, in an attempt to minimise the cushioning advantages of different shoes, the investigators in the previous study⁸ provided standardised shoes for all of their subjects. It is more clinically practical, however, to have the subjects perform the exercise with bare feet, as we did, to ensure that the same amount of work with each lift is performed on any given trial. It is also important to note that both the previous investigators and ourselves used a standard heel rise height. Theoretically, this means that people with longer feet do not have to rotate their ankles as much to reach this height, and that comparisons across individuals would not be valid. We acknowledged this limitation before our study but found that measuring the angular displacement and then setting the instrument for the proper height across a number of individuals did not in practice change the setting of the instrument. This is because the relevant distance is between the toes and the navicular, which varies very little between individuals of normal height. In addition, the test is meant as a repeated measures test where the clinician measures changes over time. Under these conditions, differences between individuals because of different foot lengths are not relevant.

This test measures the ability to raise the heel, which theoretically includes the gastrocnemius and soleus, and also muscles that extend into the foot (flexor hallucis longus, posterior tibialis, peroneus brevis, and peroneus longus). We believe that we are mostly testing the gastrocnemius and soleus, because patients with Achilles' tendon rupture cannot lift their heel off the ground. With respect to differentiating between the gastrocnemius (35–60% type I fibres) and soleus muscles (70–100% type I fibres),²⁰ the gastrocnemius is said to contribute more to the action if the knee is straight (as in our protocol) than if it is bent,²¹ but we have chosen to report the test as "triceps surae" because we did not test the differential use of the muscles in our protocol (for example by electromyographic analysis).

Conclusions

The good reliability, low cost, and simplicity of our fatigue measuring protocol make it a useful and practical measure of triceps surae muscle fatigability in subjects with no current muscle injury. Reliability in patients with a very recent triceps surae muscle strain or other medical conditions (for example, osteoarthritis) remains to be established.

ACKNOWLEDGEMENTS

The study was supported by an unrestricted grant-in-aid from the Beiersdorf-Jobst research programme of the American College of Phlebology. Michael Haber, Laurent Azoulay, and Eyal Golan were also supported by the Lady Davis Institute student challenge summer studentship 1999 (MH and LA) and the McGill Faculty of Medicine student research bursary (EG). Drs Kahn and Shrier are Chercheurs Boursiers Cliniciens (clinical research scholars) supported by the Fonds de Recherche en Santé du Québec.

Authors' affiliations

M Haber, E Golan, L Azoulay, S R Kahn, I Shrier, SMBD-Jewish General Hospital, Montreal, Quebec, Canada

REFERENCES

- 1 Croisier JL, Forthomme B, Namurois MH, *et al.* Hamstring muscle strain recurrence and strength performance disorders. *Am J Sports Med* 2002;**30**:199–203.
- 2 Garrett WE, Safran MR, Seaber AV, *et al.* Biomechanical comparison of stimulated and nonstimulated skeletal muscle pulled to failure. *Am J Sports Med* 1987;**15**:448–54.
- 3 Liemohn W. Factors related to hamstring strains. *J Sports Med* 1978;**18**:71–6.
- 4 Lindstrom L, Kadefors R, Petersen I. An electromyographic index for localized muscle fatigue. *J Appl Physiol* 1977;**43**:750–4.
- 5 Roussos C, Macklem PT. Inspiratory muscle fatigue. *Handbook of physiology – the respiratory system III*. Baltimore: Williams and Wilkins, 1986:511–27.
- 6 Bellemare F, Bigland-Ritchie B. Assessment of human diaphragm strength and activation using phrenic nerve stimulation. *Respir Physiol* 1984;**58**:263–77.
- 7 Sapega A. Muscle performance evaluation in orthopaedic practice. *J Bone Joint Surg [Am]* 1990;**72A**:1562–74.
- 8 Svantesson U, Osterberg U, Thomee R, *et al.* Muscle fatigue in a standing heel-rise test. *Scand J Rehab Med* 1998;**30**:67–72.
- 9 Edwards RHT. Physiological analysis of skeletal muscle weakness and fatigue. *Clin Sci Mol Med* 1978;**54**:463–70.
- 10 Respiratory Muscle Fatigue Workshop Group. Respiratory muscle fatigue: report of the Respiratory Muscle Fatigue Workshop Group. *Am Rev Respir Dis* 1990;**142**:474–80.
- 11 Kahn SR, Azoulay L, Hirsch A, *et al.* Acute effects of exercise in patients with previous deep venous thrombosis: impact of the postthrombotic syndrome. *Chest* 2003;**123**:399–405.
- 12 Atkinson G, Reilly T. Circadian variation in sports performance. *Sports Med* 1996;**21**:292–312.
- 13 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;**26**:217–38.
- 14 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
- 15 Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;**30**:1–15.
- 16 Atkinson G, Nevill A. Typical error versus limits of agreement. *Sports Med* 2000;**30**:375–7.
- 17 Hopkins WG. Typical error versus limits of agreement. *Sports Med* 2000;**30**:378–81.
- 18 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.
- 19 Streiner DL, Norman GR. Reliability. *Health measurement scales*. New York: Oxford University Press, 1996:104–27.
- 20 Johnson MA, Polgar J, Weightman D, *et al.* Data on the distribution of fibre types in thirty-six human muscles. An autopsy study. *J Neurol Sci* 1973;**18**:111–29.
- 21 Kendall FP, McCreary EK. *Muscle testing and function*. Baltimore: Williams and Wilkins, 2002.