Title: Short- and long-term reliability of leg extensor power measurement in middle-aged and older adults

Running title: Reliability of leg extensor power measurement

Christopher Hurst[1], Alan M Batterham[2], Kathryn L Weston[3] and Matthew Weston[4]

1. Department of Sport and Exercise Sciences, School of Social Sciences, Business and Law, Teesside University, Middlesbrough, TS1 3BA, UK.
2. Health and Social Care Institute, Teesside University, Middlesbrough, TS1 3BA, UK. a.batterham@tees.ac.uk; 01642 342771
3. Health and Social Care Institute, Teesside University, Middlesbrough, TS1 3BA, UK. k.weston@tees.ac.uk; 01642 342939
4. Department of Sport and Exercise Sciences, School of Social Sciences, Business and Law, Teesside University, Middlesbrough, TS1 3BA, UK. m.weston@tees.ac.uk; 01642 384430

Corresponding author:
Christopher Hurst
Department of Sport and Exercise Sciences,
School of Social Sciences, Business and Law,
Teesside University,
Middlesbrough,
TS1 3BA,
UK.
Email: c.hurst@tees.ac.uk
Telephone: +44 (0)1642 342308
Fax: +44 (0)1642 342399

Acknowledgements: None declared

Title: Short- and long-term reliability of leg extensor power measurement in middle-aged and older adults

Running title: Reliability of leg extensor power measurement

Abstract

Muscular power is important for maintaining physical functioning with aging. Proper quantification of the reliability of muscular power tests is crucial to inform monitoring of individuals and sample size planning for interventional studies. This study evaluated short- and long-term reliability of leg extensor power measurement in 72 adults (age 62.7 ± 8.6 years). Participants completed four repeat trials on the Nottingham leg extensor power rig, with a further trial twelve weeks later. Mean change, typical error, and intraclass correlation coefficients (ICC) were calculated. For short-term reliability, mean change in power output was trivial after two trials (1.2%-4.8%). Typical errors were small following four trials in the dominant leg of males (10.9% to 5.8%), three in the non-dominant leg of males (9.9% to 6.2%) and the dominant leg of females (10.0% to 9.6%) and two in the non-dominant leg in females (8.3%). Intraclass correlation coefficients (ICCs) were very high (0.88-0.96). For long-term reliability, mean change remained trivial (1%-2.5%), typical errors remained small (5.8-8.6%), and ICCs very high (0.94-0.96). The leg extensor power rig is a reliable method for assessing lower body muscular power, both short- and long-term, with only minimal habituation effects.

Keywords: Muscle power; functional performance; physical function; aging.

Word Count: 4020

Introduction

Age-related changes in the structure and function of the musculoskeletal system affect the ability of older adults to carry out the everyday tasks of daily living (Aagard, Suetta, Caserotti, Magnusson & Kjaer, 2010; Janssen, Heymsfield, & Ross, 2002; Skelton, Greig, Davies, & Young, 1994). Although considerable evidence has highlighted the age-related decline in skeletal muscle mass and muscular strength (Janssen, Heymsfield, Wang, & Ross, 2000; Skelton et al., 1994), muscular power may be a more important determinant of effective physical functioning in older adults (Foldvari et al., 2000; Reid & Fielding, 2012). For example, muscular power is more important than strength for activities such as chair rising and stair climbing (Bassey et al., 1992), is related to clinically important improvements in gait speed and other measures of functional performance (Bean et al., 2010; Tiggemann et al., 2016), and also plays a role in fall prevention (Skelton, Kennedy, & Rutherford, 2002).

Given the importance of maintaining lower body muscular power in older adults, its assessment is an important tool for practitioners wishing to monitor and evaluate functional capacity, as well as for researchers intending to quantify training intervention outcomes. One device used for evaluating functional lower body muscle power in older adults is the leg extensor power rig (Medical Engineering Unit, University of Nottingham, Nottingham, UK). This equipment provides a functional method for the assessment of leg extensor power (Bassey & Short, 1990; Bassey et al., 1992) employing similar muscle groups and joint angles to those used in activities such as stair climbing and rising from a chair (Bassey & Short, 1990). Lower-body muscle power measured using the leg extensor power

rig has been shown to be a predictor of physical function in older adults (Straight, Brady, & Evans, 2015a; Straight, Brady & Evans, 2015b), while the non-impact nature of the movement and simplicity of the test itself make it an appealing measurement tool.

To enable informed decision making about the appropriateness of a test, practitioners and researchers require an understanding of its reliability (Atkinson & Nevill, 1998), as high test reliability facilitates the quantification of changes that are small, yet could be practically important (Hopkins, 2015). The reliability of the performance of a test refers to the consistency or reproducibility of performance when the test is performed repeatedly (Hopkins, Schabort, & Hawley, 2001) and a statistic that captures the variability in repeated testing is the typical error (Hopkins, 2000). For many measurements in sports science and medicine, the typical error is best expressed as a coefficient of variation (CV; percentage of the mean) (Hopkins, 2000). To date, reliability evaluations of the leg extensor power rig have focused on CVs calculated over two trials, usually performed one week apart (short-term). In the original investigation of the leg extensor power rig, a CV of 9.4% in 46 participants (age range: 20-86 years) across two trials (a test-retest design), was reported (Bassey & Short, 1990). More recent investigations into the leg extensor power rig in older men (n =55, mean age 73 years [Blackwell, Cawthon, Marshall, & Brand, 2009]; n =73 men, age range 60-87 years [Schroeder et al., 2007]) and older women (n=35, aged >65 years [Skelton et al., 2002]), have reported test-retest CVs, calculated using different methods, ranging from <8% to 15.5%. However, to inform sample size estimation for an intervention using the external power output produced on the leg extensor power rig as a primary

outcome measure, reliability over the same time period as the planned intervention (long-term) should be determined (Hopkins, 2000).

While the CV represents the random variation in a measure when assessed multiple times (Hopkins, 2000), to fully quantify reliability there is also a need to consider both random error and systematic bias (Atkinson & Nevill, 1998; Batterham & George, 2000; Hopkins, 2000). Systematic bias represents the non-random error in the repeat performance of a test, whereby participants generally perform better or worse in repeat trials (Batterham & George, 2000). The change in the mean between trials provides an indication of systematic bias, which can be attributable to factors such as learning effects due to habituation, fatigue or motivation (Hopkins, 2000). Indeed, habituation is common with sports performance tests (Hopkins, 2015). The previously reported leg extensor power rig reliability data were obtained from two trials, yet this may be insufficient to eliminate systematic bias and stabilise random variability.

A full examination of any systematic bias in a measurement coupled with practical recommendations for future researchers on the number of pretest habituation sessions to employ is recommended (Atkinson & Nevill, 1998). However, reliability studies in which 50 or more volunteers perform three or more repeat trials are rare in the literature (Hopkins, 2000). As such, our aim in this study was to rigorously evaluate the short- and long-term reliability of the leg extensor power rig in a large sample of older adults tested on multiple occasions.

Method

Participants

A total of 72 community-dwelling adults aged 50-83 years took part in this investigation. Participants were physically active but were not currently, and had not in the previous year, engaged in structured exercise more than twice per week. Participants were recruited via word of mouth and advertisement at local fitness clubs, community groups and local offices. Prior to enrolment, all participants completed a medical screening questionnaire to identify any medical issues that could affect their ability to perform the required exercise. Participants with pre-existing, lower-body musculoskeletal complaints or systemic disease (e.g. diabetes mellitus, cancer, heart disease) were excluded. Following initial screening participants were excluded because of pre-existing musculoskeletal problems (n = 8) and engaging in structured exercise training (n=3). Six participants withdrew citing lack of time while one participant withdrew during the investigation because of an injury unrelated to the study. The final sample consisted of 38 males (age: 62.5 ± 8.2 years; height: 175.0 ± 5.6 cm; body mass: 88.4 ± 13.8 kg) and 34 females (age: 62.8 ± 9.2 years; height: 162.5 ± 6.0 cm; body mass: 73.1 ± 15.9 kg). All individual subjects provided written, informed consent to participate in the study, which conformed to the requirements of The Declaration of Helsinki and was approved by Teesside University Research and Ethics Committee.

Experimental procedures

To examine the reliability of the leg extensor power rig all participants completed five trials. To evaluate short-term reliability, participants were tested on four

occasions ~72h apart (Trials 1-4), as at least four trials are needed typically to properly assess habituation effects in laboratory tests (Hopkins, 2015). After these four trials, all participants returned 12 weeks later to complete a fifth trial (Trial 5) for the evaluation of long-term reliability (Figure 1). Participants were assessed on dominant and non-dominant legs separately with the dominant leg determined via a modified version of the lateral preference inventory (Coren, 1993). Testing was performed in a randomised, counterbalanced order with participants performing all testing sessions in the same order (e.g., always dominant leg first or always non-dominant leg first based on initial randomisation). All testing was performed at the same time of the day to minimise the impact of circadian variation on leg extensor power (Atkinson & Reilly, 1996) and participants were asked to avoid strenuous physical activity and alcohol in the 24 h prior to each testing session. During the intervening 12-week period, participants were instructed to maintain their habitual physical activity and not engage in any additional structured exercise.

For body and seat positioning on the power rig, we followed the testing procedures described by Bassey and Short (1990). Briefly, participants were seated with a flexed knee in an upright position with arms folded across the chest. Participants performed a unilateral leg extension until the footplate was fully depressed while the free foot rested on the floor. Seat position was determined so that the leg reached full extension at the end of the footplate movement (0.165 m). Seat position was recorded to ensure standardisation across all trials. Participants were asked to wear flat, comfortable lace-up shoes and to wear the same footwear during each trial. Participants completed a standardised warm-up prior to the testing

protocol, which consisted of three warm-up leg extensions at increasing submaximal intensity (~50, ~75 and ~90% of self-perceived maximal effort). Following completion of the warm-up, participants performed the first leg extension within 45 s. Ten maximal effort leg extensions, each separated by 30 s of passive rest were performed with participants asked to extend their leg "as hard and as fast as possible" each time. After assessment of the first leg, participants then performed the same standardised warm up as previously described, followed by the testing protocol on the second leg. The highest value recorded over the ten leg extensions was taken as the subject's peak power output for data analysis. Strong verbal encouragement was provided throughout and the first author supervised all testing sessions.

Statistical analysis

Analyses were stratified by sex and lateral preference (dominant/ non-dominant leg). Descriptive statistics were calculated for peak power and reported as mean ± SD. All analyses were performed on log-transformed data to reduce the effect of non-uniformity of error. The custom-made reliability spreadsheet of Hopkins (2015) was used, as this spreadsheet provides pairwise analyses of consecutive trials (Trial 1 v Trial 2, Trial 2 v Trial 3, Trial 3 v Trial 4, Trial 4 v Trial 5) to properly assess habituation and measurement reliability. Inferences for the between-trial changes in percentage peak power output were subsequently based on standardised thresholds for trivial, small and moderate differences of <0.2, 0.2 and 0.6 of the pooled between-subject standard deviations (Hopkins, Marshall, Batterham, & Hanin, 2009). Here, the range of thresholds for small and moderate were 5.5% to 6.8% and 17.4 to 21.7%, respectively. Typical errors were calculated

via the same reliability spreadsheet (Hopkins, 2015) and expressed as a percentage. To assess the magnitude of the typical errors, the previously described thresholds for assessing standardised mean changes were halved (<0.1, 0.1 and 0.3) (Atkinson & Batterham, 2015; Smith & Hopkins, 2011). Here, the range of thresholds for small and moderate were 2.8% to 3.4% and 8.7% to 10.9%, respectively. Between-trial reductions in typical error were considered meaningful when they crossed a magnitude threshold (e.g. 'moderate' to 'small'). The intraclass correlation coefficient ($ICC_{3,1}$; Shrout & Fleiss, 1979) was calculated (SPSS v.21, Armonk, NY: IBM Corp) with qualitative inference based on the following thresholds: >0.99, extremely high; 0.99-0.90, very high; 0.75-0.90, high; 0.50-0.75, moderate; 0.20-0.50, low; <0.20, very low (Malcata, Vandenbogaerde, & Hopkins, 2014). Uncertainty in estimates is shown as 90% confidence intervals throughout.

Results

Descriptive data for peak power across all five trials are presented in Figure 2.

Short-term reliability

Between-trial pairwise analyses and intraclass correlation coefficients are presented in Table 1. The mean change in peak power output for the dominant and non-dominant leg of males and females was trivial (1.2%-4.8%) after two trials (Trial 1 and Trial 2) and remained trivial (1.9%-5.3%) after a further two trials (Trial 3 and Trial 4). Intraclass correlation coefficients were very high for all

comparisons (0.88-0.96). Between-trial typical errors derived from the pairwise analyses of consecutive trials are presented in Figure 3. Here, the magnitude of the typical errors reduced from moderate to small following four trials in the dominant leg of males (5.8%) (Figure 3a) and following three trials in the non-dominant leg of males (6.2%) (Figure 3b) and the dominant leg in females (9.6%) (Figure 3c). The typical error was small (8.3%) after only two trials in the non-dominant leg of females (Figure 3d).

Long-term reliability

The mean change in power output between Trial 4 and Trial 5 remained trivial for the dominant and non-dominant leg of males and females (1.0%-2.5%) (Table 1). All intraclass correlation coefficients were again rated as very high (0.94-0.96) (Table 1) and all typical errors small (5.8%-8.6%) (Figure 3).

Discussion

A reliability study may be best planned to have multiple retests (Atkinson & Nevill, 1998). Through repeated tests performed on a relatively large group of middle-aged and older participants, the overarching aim of our study was to perform a rigorous evaluation of the short- and long-term reliability of the leg extensor power rig. Overall, our results demonstrate that the leg extensor power rig has good reliability for the assessment of leg power in older adults with minimal habituation effects. However, reductions in the typical error, from moderate to small, were evident with further repeat tests. Four repeat trials

performed in the short-term were sufficient for our measures of reliability to remain stable in the long-term.

Previous investigations into the reliability of the leg extensor power rig have focused only on short-term (usually one week) reliability performed over two trials (Bassey & Short, 1990; Blackwell et al., 2009; Schroeder et al., 2007; Skelton et al., 2002). The data presented here, however, provide a more detailed evaluation of reliability, as we examined the change in performance over four repeat trials; this approach is needed to properly assess habituation (Atkinson & Nevill, 1998; Hopkins, 2015). We found that in males and females, all between-trial changes in power output were trivial after only two trials and remained trivial with a further two repeat trials, suggesting minimal habituation is associated with the leg extensor power rig.

A statistical comparison of the change in the means between repeat tests should not be employed in isolation as an assessment of reliability given that very large random individual differences may still be evident when a mean change is negligible (Atkinson & Nevill, 1998). Changes in typical error over multiple repeat tests should also be appraised for researchers to make a truly informed decision over the number of pretests to employ. Indeed, it is the typical error that influences the precision of measurements in an experimental study (Hopkins, 2000). Further, a general advantage of the typical error over other indicators of reliability is that it enables extrapolation of the results of absolute reliability studies to new individuals and to compare reliability between different measurement tools (Atkinson & Nevill, 1998). In our study, performing two trials resulted in typical

errors that were classified as moderate (~10%) for the dominant leg of males and females and the non-dominant leg of males, and small (8.3%) for the non-dominant leg in females. These typical errors lie within the range of those previously described for closed-chain ergometer-based assessments of isokinetic power (Hopkins et al., 2001) and also for the leg extensor power rig itself (6-16%) (Bassey & Short, 1990; Bassey et al., 1992; Blackwell et al., 2009; Lamb, Morse, & Evans, 1995; Robertson, Frost, Doll, & O'Connor, 1998; Schroeder et al., 2007; Skelton et al., 2002).

A novel aspect of our reliability study, however, was that we were able to demonstrate for the first time that the use of further repeat trials reduces the typical error. In the non-dominant leg of males and in the dominant leg of females, typical error was reduced from moderate to small after three trials, whereas four trials were required to reduce typical error from moderate to small in the dominant leg of males. Our reductions in typical error are consistent with Hopkins and colleagues (2001) who, when examining the reliability of power in physical performance tests, reported the typical error (CV) between the first two trials to be 1.3 times greater than the CV between subsequent trials. While heterogeneous study populations combined with methodological inconsistencies make it difficult to draw comprehensive between-study comparisons, CVs after four trials (5.8% - 7.5%) are lower than those reported by Bassey & Short, (1990), Blackwell et al., (2009), Schroeder et al., (2007), Bassey et al., (1992), Robertson et al., (1998) and similar to those reported by Skelton et al. (2002) and Lamb et al. (1995).

This is the first study to determine the long-term reliability of the leg extensor power rig. This information is an important consideration for researchers and practitioners, who require an awareness of the random error associated with a measure over an equivalent time period to an intervention (Atkinson & Nevill, 1998). In the current study, change in the mean remained trivial, with a small typical error and very high ICCs in both men and women, suggesting that after four repeat baseline trials 12-week reliability is good. Comparison of long-term reliability of muscle power assessment is challenging because of a lack of current available data. However, a study from Ditroilo, Forte, McKeown, Boreham & De Vito (2011) evaluated the inter-session reliability of vertical jump performance interspersed by 4-weeks and reported CVs ranging from 2.9%-7.2% and 3.4-10.8% in middle-aged and older adults, respectively. In general, reliability is lower for longer time between trials (Hopkins, 2015); however, our results contrast this tendency and also the findings of a previous study examining both short-term and long-term reliability of a repeated sprint test in soccer players (Impellizzeri et al., 2008). Here, the authors reported a slightly greater CV in a long-term reliability study than that obtained in the short-term reliability study and stated that this was expected since as in the short-term it can be assumed that there is no true change in individuals' measurements between trials (Impellizzeri et al., 2008). It is plausible that in the present study performing multiple baseline trials helped to secure good long-term reliability.

Reliability over the same time period as the intervention is required to inform sample size estimation for an intervention using testing equipment such as the leg extensor power rig as a primary outcome measure (Hopkins, 2000). As such, our

data can be used for sample size planning for a future trial. For example, consider a 2-group randomised controlled trial (RCT) with measures of power output before and after a 12-week intervention, with a smallest worthwhile effect in the non-dominant leg of females of a change of 5.8% and a typical error observed between Trials 4 and 5 of 5.8%. For a desired precision of a 95% confidence interval width of ± 3.5 percentage points around the mean effect, the required sample size would be 43 participants per group. This sample size provides 91% power to detect a difference between groups of a change of 5.8%, with 2P=0.05. This remarkably efficient sample size for a definitive RCT powered to detect a small effect size is due to the very high 12-week reliability for this outcome measure. With a more typical correlation for objective outcome measures over a 12-week period of, say, 0.8 (vs. the observed 0.96) the required sample size would be around 200 per group.

As well as sample size determination, our reliability data can also be used for the assessment of change when monitoring an individual (Hopkins, 2000). Here, using the data described above as an example (typical error and smallest worthwhile effect both 5.8%) an individual's power output would have to increase by >11.4% to be classified as "likely to have improved" by ≥ the smallest worthwhile effect (where "likely" is defined as a probability ≥75%; Hopkins, 2004).

An important consideration when determining measurement reliability is that the time available to perform repeated tests may not be exhaustive and as such there has to be a trade-off between measurement stability and the time needed to complete testing (Ehrenbrusthoff et al., 2016). Our data show that, with the

exception of the dominant leg of males, no further meaningful reductions in typical error (i.e. moderate to small) were evident after three repeat trials. Consequently, we believe that the practical implications of our findings are that when using the leg extensor power rig, three repeat trials provide an optimal balance between measurement stability and time spent testing.

When reporting reliability data, a distinction between absolute and relative reliability should be made (Impellizzeri & Marcora, 2009). The coefficient of variation represents a measure of absolute reliability (the degree to which repeated measurements vary for individuals) (Atkinson & Nevill, 1998), whereas relative reliability (the degree to which individuals maintain their position in a sample of repeated measurements) is usually assessed via correlation coefficients (Batterham & George, 2000). In our study, the relative reliability of the leg extensor power rig showed very high ICCs for all measures. Ideally, a confidence interval for the ICC should be calculated and reported to indicate the likely range of values containing the true population ICC (Batterham & George, 2000). In this instance, our likely range for relative reliability remained high to very high for all measures.

Although similar, the results reported for dominant and non-dominant legs were not identical in this investigation with differences in change of the mean and typical errors evident between legs in both males and females. Previously, researchers have assessed right leg and left leg rather than considering leg dominance when using the leg extensor power rig; however, the results of this investigation have shown that reliability can vary between dominant and non-dominant limbs. Future investigations are encouraged to follow our approach of

considering dominant and non-dominant limbs to enable meaningful comparisons within and between studies. Researchers should categorise limbs as ‘affected’ and ‘unaffected’ when evaluating participants with skeletal or neuromuscular contraindications as per Robertson et al. (1998).

Although there is a wide range of assessment tools available for monitoring lower body muscular power there remains no consensus on the most appropriate method. Isokinetic dynamometers are often used to assess muscle power, yet these don’t reflect the real-world nature of muscular work where muscles have to overcome fixed resistances at varying velocities (Harridge, Pearson & Young, 1999). Sit-to-stand tests, of which there are a number of derivatives, are often used as a measure of lower limb strength or power (Cheng et al., 2014; McCarthy, Horvat, Holtsberg, & Wisenbaker, 2004). While evidence suggests that these are reliable tests in older populations (Jones, Rikli, & Beam, 1999), sit-to-stand performance is influenced by a number of physiological and psychological processes (Lord, Murray, Chapman, Munro, & Tiedemann, 2002) thus suggesting it may be a composite measure of a number of components of performance. Additionally, previous work has identified that both leg extensor power and standing balance are related to chair rise time supporting the idea that this is not purely a proxy measure of leg power (Hardy et al., 2010). Conversely, the leg extensor power rig provides an accessible and functionally relevant isolated assessment of lower body muscular power in a single explosive movement (Bassey et al., 1990). The leg extensor power rig can predict physical function in older adults (Straight et al., 2015a; Straight et al., 2015b), identify differences between young and older adults, fallers and non-fallers (Perry, Carville, Smith, Rutherford, & Newham, 2007) and detect training

induced improvements (Capodaglio et al., 2005; Caserotti, Aagaard, Buttrup Larsen, & Puggaard, 2008). The leg extensor power rig is also an important tool in the evaluation of pre-surgical exercise interventions with functional performance becoming an increasingly important variable of interest for clinicians (Jensen, Laustsen, Jensen, Borre, & Petersen, 2016). The findings from this study, combined with prior studies, suggest the leg extensor power rig is a reliable tool for assessing functional lower body muscular power.

The results of the present investigation are representative of a healthy population aged 50-83 years and therefore should not be extrapolated to represent all middle-aged and older adults. Consequently, further investigation is needed to understand the long-term reliability of the leg extensor power rig in older participants and in populations with musculoskeletal complications. Further study should also evaluate long-term reliability of other methods of leg power assessment so that meaningful comparisons can be made.

Conclusion

Our investigation is the first to evaluate both short- and long-term reliability of the leg extensor power rig, with our findings suggesting that this is a reliable method for assessing leg extension power in both the short- and long-term with only minimal habituation effects. However, performing repeated tests reduces typical errors from moderate to small, with the number of pretests required varying between males and females, and dominant and non-dominant legs. For researchers using the leg extensor power rig as an outcome measure in an intervention study, our data suggest that performing three repeat trials will provide an appropriate

balance between measurement stability and the time demands of testing. In a wider context, researchers are encouraged to evaluate short- and long-term reliability of their outcome measures to best inform the planning and delivery of future intervention studies.

Conflict of Interest The authors declare that they have no competing interests to report.

References

Aagaard, P., Suetta, C., Caserotti, P., Magnusson, S.P., & Kjaer, M. (2010). Role of the nervous system in sarcopenia and muscle atrophy with aging: strength training as a countermeasure. *Scandinavian Journal of Medicine and Science in Sports*, 20(1), 49-64. doi: 10.1111/j.1600-0838.2009.01084.x

Atkinson, G., & Batterham, A.M. (2015). True and false interindividual differences in the physiological response to an intervention. *Experimental Physiology*, 100(6), 577-588. doi: 10.1113/EP085070

Atkinson, G., & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26(4), 217-238. doi: 10.2165/00007256-199826040-00002

Atkinson, G., & Reilly, T. (1996). Circadian variation in sports performance. *Sports Medicine*, 21(4), 292-312. doi: 10.2165/00007256-199621040-00005

Bassey, E.J., Fiatarone, M.A., O'Neill, E.F., Kelly, M., Evans, W.J., & Lipsitz, L.A. (1992). Leg extensor power and functional performance in very old men and women. *Clinical Science*, 82(3), 321-327. doi: 10.1042/cs0820321

Bassey, E.J., & Short, A.H. (1990). A new method for measuring power output in a single leg extension: feasibility, reliability and validity. *European Journal of Applied Physiology*, 60(5), 385-390. doi: 10.1007/BF00713504

Batterham, A.M., & George, K.P. (2000). Reliability in evidence-based clinical practice: a primer for allied health professionals. *Physical Therapy in Sport*, 1(2), 54-62. doi:10.1054/ptsp.2000.0010

Bean, J.F., Kiely, D.K., LaRose, S., Goldstein, R., Frontera, W.R., & Leveille, S.G. (2010). Are Changes in Leg Power Responsible for Clinically Meaningful

Improvements in Mobility in Older Adults? *Journal of the American Geriatrics Society*, 58 (12), 2363-2368. doi: 10.1111/j.1532-5415.2010.03155.x

Blackwell, T., Cawthon, P.M., Marshall, L.M., & Brand, R. (2009) Consistency of Leg Extension Power Assessments in Older Men. *American Journal of Physical Medicine & Rehabilitation*, 88(11), 934-940. doi: 10.1097/PHM.0b013e3181bbddfb

Capodaglio, P., Capodaglio, E.M., Ferri, A., Scaglioni, G., Marchi, A., Saibene, F. (2005). Muscle function and functional ability improves more in community-dwelling older women with a mixed-strength training programme. *Age and Ageing*, 34(2), 141-147. doi: 10.1093/ageing/afi050

Caserotti, P., Aagaard, P., Buttrup Larsen, J., & Puggaard, L. (2008). Explosive heavy-resistance training in old and very old adults: changes in rapid muscle force, strength and power. *Scandinavian Journal of Medicine and Science in Sports*, 18(6),773-782. doi: 10.1111/j.1600-0838.2007.00732.x

Cheng, Y-Y., Wei, S-H., Chen, P-Y., Tsai, M-W., Cheng, L-C., Liu, D-H., & Kao C-L. (2014) Can sit-to-stand lower limb muscle power predict fall status? *Gait & Posture*, 40(3), 403-407. doi: 10.1016/j.gaitpost.2014.05.064

Coren, S. (1993). The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness: Norms for young adults. *Bulletin of the*

Psychonomic Society, 31(1), 1-3. doi: 10.3758/BF03334122

Ditroilo, M., Forte, R., McKeown, D., Boreham, C., & De Vito, G. (2011). Intra- and inter-session reliability of vertical jump performance in healthy middle-aged and older men and women. *Journal of Sports Sciences*, 29 (15), 1675-1682. doi: 10.1080/02640414.2011.614270

Ehrenbrusthoff, K., Ryan, C.G., Gruneberg, C., Wolf, U., Krenz, D., Atkinson, G., & Martin, D. (2016). The intra- and inter-observer reliability of a novel protocol for two-point discrimination in individuals with chronic low back pain. *Physiological Measurement*, 37(7), 1074-1088. doi: 10.1088/0967-3334/37/7/1074

Foldvari, M., Clark, M., Laviolette, L.C., Bernstein, M.A., Kaliton, D., Castaneda, C., Pu, C.T., Hausdorff, J.M., Fielding, R.A, & Fiatrone Singh, M.A. (2000). Association of muscle power with functional status in community-dwelling elderly women. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 55(4), M192-199. doi: 10.1093/gerona/55.4.M192

Hardy, R., Cooper, R., Shah, I., Harridge, S., Guralnik, J., & Kuh, D. (2010). Is chair rise performance a useful measure of leg power? *Aging Clinical and Experimental Research*, 22(5-6), 412-418. doi: 10.1007/BF03324942

Harridge, S.D.R., Pearson, S.J., & Young, A. (1999). Muscle power loss in old age: functional relevance and effects of training. In Capodaglio, P., & Narici, M.V.

(Eds.) *Physical activity in the elderly* (pp 123-137). Maugeri Foundation Books and PI-ME Press, Pavia.

Hopkins, W.G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30(1), 1-15. doi: 10.2165/00007256-200030010-00001

Hopkins, W.G. (2004). How to interpret changes in an athletic performance test. *Sportscience*, 8, 1-7. (sportsci.org/jour/04/wghtests.htm)

Hopkins, W.G. (2015). Spreadsheets for analysis of validity and reliability. *Sportscience*, 19, 36-42. (sportsci.org/2015/ValidRely.htm)

Hopkins, W.G., Marshall, S.W., Batterham, A.M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise*, 41(1), 3-13. doi: 10.1249/MSS.0b013e31818cb278

Hopkins, W.G., Schabort, E.J., & Hawley, J.A. (2001). Reliability of power in physical performance tests. *Sports Medicine*, 31(3), 211-234. doi: 10.2165/00007256-200131030-00005

Impellizzeri, F.M., & Marcora, S.M. (2009). Test validation in sport physiology: lessons learned from clinimetrics. *International Journal of Sports Physiology and Performance*, 4(2), 269-277.

Impellizzeri, F.M., Rampinini, E., Castagna, C., Bishop, D., Ferrari Bravo, D.,

Tibaudi, A., & Wisloff, U. (2008). Validity of a repeated-sprint test for football. *International Journal of Sports Medicine*, 29(11), 899-905. doi: 10.1055/s-2008-1038491

Janssen, I., Heymsfield, S.B., & Ross, R. (2002). Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability. *Journal of the American Geriatrics Society*, 50(5), 889-896. doi: 10.1046/j.1532-5415.2002.50216.x

Janssen, I., Heymsfield, S.B., Wang, Z.M., & Ross, R. (2000). Skeletal muscle mass and distribution in 468 men and women aged 18-88 yr. *Journal of Applied Physiology*, 89(1), 81-88.

Jensen, B.T., Laustsen, S., Jensen, J.B., Borre, M., & Petersen, A.K. (2016). Exercise-based pre-habilitation is feasible and effective in radical cystectomy pathways—secondary results from a randomized controlled trial. *Supportive Care in Cancer*, 24(8), 3325-3331. doi: 10.1007/s00520-016-3140-3

Jones, C.J., Rikli, R.E., & Beam, W.C. (1999). A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Research Quarterly for Exercise and Sport*, 70(2), 113-119. doi: 10.1080/02701367.1999.10608028

Lamb, S.E., Morse, R.E., & Evans, J.G. (1995). Mobility after proximal femoral fracture: the relevance of leg extensor power, postural sway and other factors. *Age and Ageing*, 24(4), 308-314. doi: 10.1093/ageing/24.4.308

Lord, S.R., Murray, S.M., Chapman, K., Munro, B., & Tiedemann, A. (2002). Sit-to-stand performance depends on sensation, speed, balance, and psychological status in addition to strength in older people. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 57(8), M539-M543. doi: 10.1093/gerona/57.8.M539

Malcata, R.M., Vandenbogaerde, T.J., & Hopkins, W.G. (2014). Using athletes" world rankings to assess countries" performance. *International Journal of Sports Physiology and Performance*, 9(1), 133-138. doi: 10.1123/ijspp.2013-0014

McCarthy, E.K., Horvat, M.A., Holtsberg, P.A., & Wisenbaker, J.M. (2004). Repeated chair stands as a measure of lower limb strength in sexagenarian women. *The Journals of Gerontology Series A: Biological Sciences and* Medical, 59(11), 1207-1212. doi: 10.1093/gerona/59.11.1207

Perry, M.C., Carville, S.F., Smith, I.C.H., Rutherford, O.M., & Newham, D.J. (2007). Strength, power output and symmetry of leg muscles: effect of age and history of falling. *European Journal of Applied Physiology*, 100(5), 553–561. doi: 10.1007/s00421-006-0247-0

Reid, K.F., & Fielding, R.A. (2012). Skeletal muscle power: a critical determinant of physical functioning in older adults. *Exercise and Sport Sciences Reviews*, 40(1), 4-12. doi: 10.1097/JES.0b013e31823b5f13

Robertson, S., Frost, H., Doll, H., & O'Connor, J.J. (1998). Leg extensor power and quadriceps strength: an assessment of repeatability in patients with osteoarthritic knees. *Clinical Rehabilitation*, 12(2), 120-126. doi: 10.1191/026921598673072472

Schroeder, E.T., Wang, Y., Castaneda-Sceppa, C., Cloutier, G., Vallejo, A.F., Kawakubo, M., Jensky, N.E., Coomber, S.M., Azen, S.P., & Sattler, F.R. (2007). Reliability of maximal voluntary muscle strength and power testing in older men. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 62(5), 543-549. doi:10.1093/gerona/62.5.543

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. doi: 10.1037/0033-2909.86.2.420

Skelton, D.A., Greig, C.A., Davies, J.M., & Young, A. (1994). Strength, power and related functional ability of healthy people aged 65-89 years. *Age and Ageing*, 23(5), 371-377. doi: 10.1093/ageing/23.5.371

Skelton, D.A., Kennedy, J., & Rutherford, O.M. (2002). Explosive power and asymmetry in leg muscle function in frequent fallers and non-fallers aged over 65. *Age and Ageing*, 31(2), 119-125. doi: 10.1093/ageing/31.2.119

Smith, T.B., & Hopkins, W.G. (2011). Variability and Predictability of Finals Times of Elite Rowers. *Medicine and Science in Sports and Exercise*, 43(11),

2155-2160. doi: 10.1249/MSS.0b013e31821d3f8e

Straight, C.R., Brady, A.O., & Evans, E.M. (2015a). Sex-specific relationships of physical activity, body composition and muscle quality with lower-extremity physical function in older men and women. *Menopause*, 22 (3), 297-303. doi: 10.1097/gme.0000000000000313.

Straight, C.R., Brady, A.O., & Evans, E.M. (2015b). Muscle quality and relative adiposity are the strongest predictors of lower-extremity physical function in older women. *Maturitas*, 80, 95-99. doi: 10.1016/j.maturitas.2014.10.006.

Tiggemann, C.L., Dias, C.P., Radaelli, R., Massa, J.C., Bortoluzzi, R., Schoenell, M.C.W., Noll, M., Alberton, C.L., & Kruel, L.F.M. (2016). Effect of traditional resistance and power training using rated perceived exertion for enhancement of muscle strength, power, and functional performance. *Age*, 38(2), 42. doi: 10.1007/s11357-016-9904-3

Table 1 Pairwise comparisons of the short- and long-term reliability for the leg extensor power rig

		Short-term reliability			**Long-term reliability**
Males (n=38)		Trial 1 v Trial 2	Trial 2 v Trial 3	Trial 3 v Trial 4	Trial 4 v Trial 5
Dominant leg	Mean change (%)	4.0	0.9	1.9	2.5
	90% CI	0 to 8.3	-2.9 to 4.9	-0.3 to 4.2	0 to 5.1
	Qualitative inference	*Trivial*	*Trivial*	*Trivial*	*Trivial*
	ICC	0.88	0.88	0.96	0.95
	90% CI	0.81 to 0.93	0.80 to 0.93	0.93 to 0.98	0.92 to 0.97
	Qualitative inference	*Very High*	*Very High*	*Very high*	*Very high*
Non-dominant leg	Mean change (%)	1.2	-1.4	2.7	1.7
	90% CI	-2.4 to 4.9	-3.6 to 1.0	0.5 to 5.0	-0.8 to 4.3
	Qualitative inference	*Trivial*	*Trivial*	*Trivial*	*Trivial*
	ICC	0.91	0.96	0.96	0.95
	90% CI	0.84 to 0.94	0.93 to 0.98	0.94 to 0.98	0.92 to 0.97
	Qualitative inference	*Very high*	*Very high*	*Very high*	*Very high*
Females (n=34)					
Dominant leg	Mean change (%)	4.8	2.4	2.1	1.2
	90% CI	0.8 to 9.0	-1.4 to 6.3	-0.6 to 4.8	-2.2 to 4.7
	Qualitative inference	*Trivial*	*Trivial*	*Trivial*	*Trivial*
	ICC	0.91	0.92	0.96	0.94
	90% CI	0.84 to 0.95	0.87 to 0.96	0.93 to 0.98	0.89 to 0.96
	Qualitative inference	*Very high*	*Very high*	*Very high*	*Very high*
Non-dominant leg	Mean change (%)	3.4	1.3	5.3	1.0
	90% CI	0.1 to 6.9	-1.6 to 4.2	2.2 to 8.5	-1.4 to 3.3
	Qualitative inference	*Trivial*	*Trivial*	*Trivial*	*Trivial*
	ICC	0.92	0.95	0.94	0.96
	90% CI	0.87 to 0.96	0.91 to 0.97	0.90 to 0.97	0.93 to 0.98
	Qualitative inference	*Very high*	*Very high*	*Very high*	*Very high*

ICC = Intraclass correlation coefficient; CI = confidence intervals

Figure Captions

Figure. 1 Schematic of study design.

Figure. 2 Peak power output (watts) from each trial. Closed squares represent short-term trials (1-4) and open diamonds represent long-term trials (5). Error bars represent SD.

Figure. 3 Short- and long-term between-trial typical errors (coefficient of variation, %) for the leg extensor power rig. (a = male, dominant leg; b = male, non-dominant leg; c = female, dominant leg; d = female, non-dominant leg). Solid horizontal lines represent 90% confidence intervals. Dashed vertical lines represent thresholds for trivial, small and moderate effect sizes for the typical error. *A change in typical error across a threshold was considered meaningful.