Title: The test–retest reliability of four functional mobility tests in apparently healthy adults

Authors: Matthew J. Northgraves[1,2]*, Stephen C. Hayes[1], Philip Marshall[1], Leigh A. Madden[3], and Rebecca V. Vince[1]

[1]Department of Sport, Health and Exercise Science, University of Hull, Hull, UK

[2]Department of Health Sciences, University of York, York, UK

[3]School of Biological, Biomedical and Environmental Sciences, University of Hull, Hull, UK

***Correspondence:**

Mr Matthew J. Northgraves

Department of Sport, Health and Exercise Science, University of Hull, Hull, HU6 7RX, UK

Telephone: +44 (0)1482 466080 Email: m.northgraves@hull.ac.uk

ABSTRACT

BACKGROUND: Simple field tests are often used to assess functional mobility in clinical settings. Despite having many benefits, these tests are susceptible to measurement error and individual variation.

OBJECTIVES: To examine the test-retest and absolute reliability of timed up and go test (TUG), five times sit-to-stand (FTSTS), stair climb test (SCT) and 6 minute walk (6MWT).

METHODS: Over two sessions, thirty-five subjects (30-74 years), repeated the five tests approximately four weeks apart. Test-retest reliability (intraclass correlations [ICC]) and absolute reliability (95% limit of agreements [95% LOA]; standard error of measurement [SEM] and minimum detectable change [MDC]) were calculated.

RESULTS: All five tests had high test-retest reliability (ICC > 0.95) although significant between session changes were present for the TUG and FTSTS ($p < 0.05$). FTSTS displayed the greatest measurement error whilst 95% LOA was the most conservative measure of absolute reliability.

CONCLUSIONS: The results of this study indicate that the TUG, FTSTS, SCT and 6MWT are reliable when performed four weeks apart. Furthermore, the inclusion of SEM, MDC and 95% LOA provides reference values to aid in identifying changes over time above those of measurement error and individual variation.

KEYWORDS: TUG, FTSTS, Stair climb, 6MWT

1. INTRODUCTION

Functional mobility is the ability of an individual to carry out everyday activities such as rising from a chair, walking to the shops or even putting on socks. As a result of ageing, declines in cardiorespiratory fitness, muscular strength and endurance, and/or a loss of balance [1, 2] can all occur, contributing to impaired functional mobility and health related quality of life in the individual [3]. Undergoing a major surgical procedure can equally have a debilitating effect on the individual with prolonged periods of immobilisation promoting acute insulin resistance, reduced body mass and muscle wasting [4]; all of which accentuate the decay in functional mobility further.

The use of functional mobility tests remain a popular metric by which to assess changes in physical functioning in both clinical and ageing populations. Various tests have been developed to assess the various components which can impact on the mobility of an individual. For example, poor performance of the timed up and go test (TUG), which is considered a measure of both balance [5] and functional mobility [6], has been associated with increased incidences of falls in elderly populations [5] whilst the 6 minute walk test (6MWT) distance has been associated with all-cause mortality in chronic heart failure patients [7]. An important aspect to these tests is that they often need only a short administration time and do not require specialist equipment making them assessable in a host of clinical settings, easy to administer and simple for the patient/client to perform. They do, however, have certain limitations as their sensitivity to change over longer periods is potentially compromised by the presence of measurement error and variation in individual performance.

An understanding of the test-retest reliability is therefore imperative in interpreting the results of each specific test. Intraclass correlation coefficients (ICC) remain one of the most frequently used statistical methods for assessing test-retest reliability [8] however these only provide a measure of relative reliability and therefore provide no indication of measurement error. As a measure of absolute reliability, the standard error of measurement (SEM) allows measurement error to be displayed in the same units as the original measurement [9]. Additionally, the minimum detectable change (MDC) can be calculated as the smallest difference between repeated trials that is not due to chance variation [10].

The aim of this study was therefore to establish the test-retest reliability and absolute reliability of four commonly used tests of functional mobility when repeated approximately four weeks apart.

2. METHODS

2.1. SUBJECTS

A sample of 35 volunteers (18 males, 17 females) was recruited from the local community via advertisement for this study. Inclusion criteria included being an apparently healthy male and female aged 30-75 years. Exclusion criteria included any history of cardiopulmonary conditions, any musculoskeletal and/or orthopaedic conditions, current injury, history of fracture within the last year, uncorrected visual impairment, recent history of dizziness or fainting, vestibular disorders and shortness of breath with minimum exertion. Participants were screened for eligibility through the completion of an institution approved pre-exercise medical questionnaire. All participants provided written informed consent, and the study was approved by the Department of Sport, Health and Exercise Human Ethics Committee and followed the principles outlined in the Declaration of Helsinki.

2.2. EXPERIMENTAL DESIGN

As the purpose of this study was to test the test-retest reliability of the four assessment measures rather than inter-rater reliability, all trials were conducted by a single tester; this ensured maximum consistency for data collection of each variable. Participants were required to attend two identical testing sessions separated by approximately four weeks. Both sessions were conducted at the same time of day in order to control for circadian variation and participants were asked to refrain from strenuous exercise in the 24 hours preceding each visit. The order of testing was the TUG, followed by the five times sit to stand (FTSTS), stair climb test (SCT) and finally the 6MWT.

1. TUG: From a plastic chair measuring 40 cm from the floor and 39 cm deep, participants were asked to stand from a seated position, walk 3 metres before turning 180° and returning to the chair to sit down. Timing started with the count of “THREE,

TWO, ONE, GO" and ended when they had returned to the seated position. Participants were instructed to perform this 'as quickly as possible but in a controlled manner' with time taken measured in seconds [5].

2. FTSTS: Using a chair as above, participants were instructed the aim of the test was to perform five sit to stand movements as fast as they could in a controlled safe manner. From an upright seated position with their back against the chair backrest and arms crossed over their chest, the test started with the count of "THREE, TWO, ONE, GO" [11].

3. SCT: Using a set of freestanding wooden stairs which consisted of five steps (each 20 cm high) and a supporting handrail, participants were required to climb to the top as quickly as possible in a controlled safe manner. The use of the handrails and walking aids was permitted if required. Participants were instructed that the tested started with the count of "THREE, TWO, ONE, GO" with the participant beginning the ascent on "GO" and the test finishing once both feet were flat on the top step [12].

4. 6MWT: A 30 metre flat walking surface was set out with cones marking each 3 metre interval with distinct markers at the start and end. Following a period of 10 minutes seated rest, participants were instructed to walk as far and as fast as possible in 6 minutes. Rest periods were permitted however time was not stopped. A standardised protocol was used in line with the guidelines provided by the ATS [13]. At the end of the 6 minutes, participants stopped when instructed with the total distance walked providing the primary outcome measure. Measures of heart rate (HR) and arterial oxygen saturation (SaO_2) (Nonin Onyx finger pulse Oximeter, Nonin Medical Inc, Plymouth, Minnesota) were taken prior to (HR_{pre}, SaO_{2pre}) and immediately after the

6MWT (HR_{post}, SaO_{2post}). Heart rate was measured at one minute intervals throughout the test allowing the average HR (HR_{ave}) to be calculated.

For TUG, FTSTS, and SCT, following an unrecorded familiarisation trial, the mean of three trials were taken for analysis. A single trial per session was performed for the 6MWT.

2.3. STATISTICAL ANALYSIS

All statistical analyses were conducted using SPSS Version 20 for windows (SPSS Inc., Chicago, Il, USA) with the exception of the Bland-Altman plots which were performed using SigmaPlot Version 12 (Systat Software, San Jose, CA, USA). Normality of data was assessed using the Shapiro-Wilks test and all data conformed to normal distribution allowing parametric statistical procedures to be used. Differences between the two testing sessions for each assessment measure were assessed using paired sample t-tests.

Relative reliability was assessed using the ICC model 3 [14]. As the mean of three trials was used for the TUG, FTSTS and SCT, test-retest reliability was measured using $ICC_{3,2}$ model. For the 6MWT, which involved a single trial each session the $ICC_{3,1}$ model was used. Absolute reliability was expressed using 95% limits of agreement (95% LOA) [15], SEM and minimum detectable change at a 95% confidence interval

(MDC_{95}). The 95% LOA represents the expected range of difference scores for each test. The SEM allowed measurement error to be displayed in the same units as the original measurement and was calculated using the formula:

$$\text{SEM} = \text{SD} \times \sqrt{(1\text{-ICC})}$$

where SD was the standard deviation for all observations from test sessions 1 and 2 and ICC was the reliability coefficient. Measurement error was also expressed as a percentage of the mean ($SEM_{\%}$) using the formula:

$$\text{SEM}_{\%} = (\text{SEM/mean}) \times 100$$

This represents the smallest change required to indicate real change in a group of participants. MDC_{95} was calculated to represent the magnitude of change required to exceed the anticipated measurement variation, measurement error and variability of participants with 95% confidence [10]. The formula used for calculating MDC_{95} was:

$$\text{MDC}_{95} = \text{SEM} \times 1.96 \times \sqrt{2}$$

where the value of 1.96 represents the 95% CI and $\sqrt{2}$ accounted for the added uncertainty in measurement associated with repeated trials. Statistical significance was set at $p \leq 0.05$ for all tests.

3. RESULTS

Thirty-five participants (18 males and 17 females; age 54.6 ± 12.1 years [Range: 30-74 years], height 170.9 ± 11.0 cm [Range: 145.6 - 195.6 cm], body mass 78.4 ± 17.8 kg

[Range: 43.0 ± 119.3 kg]) were recruited to this study. The mean number of days between trials was 27.9 ± 1.5 days [Range: 24 – 33 days]. Thirty one (17 males and 14 females) of the 35 participants reported their self-reported physical activity level as either moderately active or active. Three participants (1 male and 2 females) were sedentary whilst one female reported their physical activity level as highly active.

3.1. TUG, FTSTS AND SCT

A mean percentage improvement in the performance time of TUG (3.4%; Range: -10.4 to +16.0%), FTSTS (3.9%; Range: +20.5 to -23.7%) and SCT (1.7%; Range: +12.4 to -0.3%) was seen between the first and second visit. The improvement however was only significant ($p < 0.05$) for the TUG and FTSTS (Table 1). The results relating to the both relative (ICC) and absolute reliability (LOA, SEM & MDC) of the TUG, FTSTS and SCT are displayed in Table 2. All three tests demonstrated good test-retest reliability with high ICCs ranging from 0.96 to 0.98. Out of the three tests, the SCT displayed the greatest absolute reliability with the SEM represented as a percentage of the mean being 2.8% whilst the FTSTS had the greatest measurement error at 5.8% of the mean.

When analysed based on gender, mean performance time for all three tests was faster in males (Table 1), however neither relative nor absolute reliability were greatly affected (Table 2). The magnitude of the ICCs for all three tests remained similar in males (ICCs = 0.97 to 0.98) and females (ICCs = 0.94 to 0.97) compared to when all participants were combined (ICCs 0.96 to 0.98). In respects to absolute reliability, the greatest variability between genders was observed in the FTSTS.

3.2. 6MWT

A mean improvement of approximately 5.6 metres (+0.9%) was seen between the first and second visit although this was not significant ($p > 0.05$) (Table 3). No significant difference was seen between sessions for SaO_{2post}, HR_{pre}, HR_{post} or HR_{ave} however SaO_{2pre} was significantly lower in session 2. The high ICC and narrow accompanying 95% CI demonstrated good test-retest reliability for the 6MWT (Table 4). Furthermore, the values reported for both 95% LOA and MDC_{95} were similar whilst the SEM of 13.7 metres ($SEM_{\%}$ -2.3%) represented a low value of measurement error.

When analysed based on gender, the mean distance walked was significantly further (+12.1 metres; +2.0%; $p < 0.05$) in the 2nd session for males however no difference between sessions was evident for females (-1.2 metres; 0.2%; $p > 0.05$). Despite the difference in males between sessions neither the relative nor absolute reliability of the 6MWT was greatly affected.

4. DISCUSSION

The aims of this study were 1). to establish the test-retest reliability of four functional mobility tests often used within clinical studies when performed approximately four weeks apart and 2). to calculate LOA, SEM and MDC, giving an indication of absolute reliability between repeated tests. All four tests used in this study displayed good test-retest reliability, exceeding the ICC threshold of 0.90 previously reported to be required for a clinical test [16]. Whilst the use of ICC provide an indication of the relative reliability of a test, the inclusion of a measure of absolute reliability is important in order to gain an understanding of whether real change has actually occurred. In the current study despite good test-retest reliability being seen for all the tests used, considerable individual performance variability was present for some tests (in particular the FTSTS), highlighting the need to incorporate both measures of relative and absolute agreement when assessing the reliability of a test [17].

Of the four tests included in the current study, the 6MWT is probably the most frequently used acting as a means of assessing the effectiveness of different intervention programmes [18] as well as a predictor of both cardiorespiratory fitness [19] and clinical outcomes [7]. As in the current study, good test-retest reliability has been observed in a number of other populations including cardiac patients (ICCs = 0.88 - 0.97) [20-22], type 2 diabetics (ICC = 0.99) [23] and the elderly (ICCs = 0.87 – 0.93) [24]. It is however often reported that at least one, if not more, familiarisation trials are required in order to alleviate any potential learning effect and thus achieve a consistent baseline measurement for the 6MWT [21, 22, 26, 27].

In healthy individuals aged 60-70; it was only from the third trial that the measurement became reliable when performing five 6MWT over a 1 week period [26]. Between both the 1st and 2nd, and 2nd and 3rd trials a mean increase of ~20 metres was reported; representing a 3.7 – 3.8% increase between trials. An average improvement of 8 ± 5% (+47 metres) in the second of two trials performed on the same day was observed in healthy individuals aged 50 - 85 years [27]. Both Hanson et al. [22] and Hamilton et al. [21] reported a learning effect occurred between trials within a cardiac rehabilitation setting despite reporting good relative reliability (ICC=0.91 and 0.97 respectively). An 11.8% (+52 metres) increase in distance walked was observed in Hanson et al. [22] between the 1st and 2nd trial and this increased to 19.1% (+85 metres) between the 1st and 3rd trial. Furthermore, whether the three tests were performed on the same day or spread over a week did not alter the presence of the learning effect [22]. Although the improvement was smaller, Hamilton et al. [21] observed a 3.5% (+18 metres) increase between the 1st and 2nd trial and 5.6% (+29 metres) between the 1st and 3rd trial.

Whilst performing repeated trials of the 6MWT on the same day has been shown to be physically tolerable in clinical populations [26, 28], it may not always be feasible. In the current study only a 0.9% (+5.6 metres) increase was witnessed between trials when all participants were combined. Even in males alone, where a 2.0% (+12.1 metres) increase in distance walked was observed during the 2nd trial compared to the 1st, the magnitude of the change was lower than some of the values previously reported [21, 26, 27]. This may indicate to a certain extent that any learning effect gained through previously performing the test may be attenuated by the longer period (4 weeks) between trials compared to those repeated over a shorter period of time (1 – 14 days) [21, 26, 27].

Furthermore, the absence of a significant difference in HR_{post}, HR_{ave} or SaO_{2post} between the sessions (Table 3) would suggest there was no increased or decreased physical effort exerted by participants during the 2^{nd} trial; potentially supporting the presence of an attenuated learning effect.

It is acknowledged that direct comparisons between this study and those using clinical populations are difficult as considerable variation does exist between population groups. The SEM (13.7 metres) and MDC_{95} (37.8 metres) seen in the current study were comparable to those reported in older type 2 diabetics (SEM = 9.88 metres; MDC_{95} = 27.37 metres) by Alfonsa-Rosa et al. [23]. This was despite only a 1 week period existing between their trials suggesting any learning effect was absent in their study [23]. These values however do differ from those seen in both elderly (SEM: 32-34 metres; MDC_{95}: 88.7-95 metres [24] and cardiac (SEM: 18.4-32.6 metres; MDC_{95}: 50.92 – 90.3 [20, 21, 29] populations therefore patient characteristics and conditions need to be considered in determining changes in performance.

Unlike with the 6MWT, the presence of a significant statistical decrease in time taken to perform the TUG and FTSTS between the first and second sessions suggested a learning effect was present. Similar improved FTSTS performance times have previously been reported in trials separated by 4-10 days [30] up to six weeks [31, 32]. Despite this, the ICC for all three studies was in excess of 0.80 indicating good correlation and agreement between trials. The ICC of 0.97 for the TUG in the current study (Table 1) exceeded that of Jette et al. [33], who reported an ICC of 0.74 in elderly frail individuals. However, the difference in study populations is likely to have influenced the reduced ICC in Jette et al. [33] compared to the current study. It is also worth noting that whilst the median number of days between trials was 14 days in Jette et al. [33], the overall range between trials varied from 0 days to 132 days. It is therefore plausible that the decrease in test-retest reliability, as indicated by ICC, was related to a true change in the study populations' ability to perform the FTSTS; especially in the individuals with the largest number of days between trials.

The results relating to the relative reliability of the FTSTS when performed with an extended period between trials have previously been varied [34]. In trials separated by 4-10 days, Bohannon et al [30] reported good test-retest reliability (ICC = 0.96; 95% CI: 0.92-0.98) in community-dwelling men and women aged 15-85 years. In contrast, when the interval between trials has been longer, lower ICC's have tended to be reported. In two studies by Schaubert and Bohannon [31, 32] in which testing sessions were separated by 6 weeks, ICCs of 0.82 (95% CI: 0.68-0.92) and 0.81 (95% CI: not stated) respectively were reported. In the current study, despite the 4 week period between tests, test-retest reliability remained good with the ICC of 0.96 far exceeding those seen in the two aforementioned studies.

This difference could potentially be explained by a number of factors, including the presence of a shorter four week period between testing sessions in the current study as opposed to six weeks [31, 32]. Furthermore, the sample sizes used in both these studies (n=21 [31] and n=11 [32]) were smaller than those of the current study (n=35). A more pertinent factor however is probably the difference in participant ages between the studies. It is acknowledged that the mean ages in both Schaubert and Bohannon studies [31, 32] (75.0 ± 5.9 years [Range: 65-85 years] and 75.5 ± 5.8 years [Range: 65-85 years] respectively) make their findings more generalizable, especially to older populations where the FTSTS is more traditionally used, than the current study (54.6 ± 12.1 years [Range: 30 -74 years]). Despite this, the current study adds to the existing literature with regards to the potential measurement error of the four tests investigated.

Whilst TUG and FTSTS displayed good relative test-retest reliability in the current study, the absolute reliability for the tests did reflect the presence of considerable individual variation in the performance of each. Inconsistencies in the agreement of relative and absolute reliability measures have previously been observed making the use of a combined approach important [17]. The FTSTS was the most variable with a $SEM_{\%}$ of 5.8% and $MDC_{95\%}$ of 16.09%. These values were less than the $SEM_{\%}$ of 6.3% and $MDC_{95\%}$ of 17.5% reported by Goldberg et al. [11] when performing repeated trials on the same day in apparently healthy older female participants. Furthermore Goldberg et al. [11] indicated a $MDC_{95\%}$ of 17.5% may be considered a low minimum change percentage. Further variation existed in the level of absolute reliability depending on the measure by which it was assessed.

The use of 95% LOA as a measure of absolute reliability in the current study reflected the most conservative method. For the FTSTS, 95% LOA suggested a change of over 2.55 seconds was required to detect real change compared to the 1.60 seconds according to the MDC_{95} (Table 2). Understanding the variation present in both the performance of the test and the different methods of calculating absolute reliability could be important when assessing any change present in repeated performances.

Although in the current study the SCT displayed good relative test-retest reliability (ICC = 0.98; 95% CI 0.95-0.99) and absolute reliability (SEM = 0.08 s; MDC_{95} = 0.22 s), the results remain difficult to interpret. Variations of the SCT have been used in a variety of different populations including those with orthopaedic limitations and the elderly. The intra-session reliability in elderly individuals (mean age 69.4 years) with hip and/or knee osteoarthritis was reported to be good with an ICC of 0.94 (95% CI 0.75-0.98) and SEM of 0.28 seconds seen for a four step ascent only SCT [12]. When performing a five step SCT including both the ascent and descent of the stairs two weeks apart, Rejeski et al. [35] reported good test- retest reliability (ICC = 0.93; 95% CI Not reported) in patients with knee osteoarthritis. Despite similar ICC being reported in Lin et al. [12], Rejeski et al. [35] and the current study, making comparisons between the studies is difficult. The absence of any limiting condition such as osteoarthritis in the present study that may have impaired the ability of participants to climb stairs, means the performance time of 2.77 seconds is faster than those reported in either Lin et al. [12] (4.17 ± 2.80 s) or Rejeski et al. [35] (10.21 ± 4.45 s). It is therefore acknowledged the SCT results are difficult to generalise beyond the present study.

This study is not without limitations. The use of an apparently healthy population with a relatively wide age range (30-74 years) in this study means the results cannot be directly generalised to those of a specific clinical population. Furthermore, given the sample size, stratification based on factors such as age, gender and self-reported physical activity was not possible. The sub-analysis based on gender alone (Tables 2 and 4) did not differ greatly between the genders for any of the tests in the current study, however whether a more pronounced difference would be observed with a larger sample size cannot be dismissed.

Despite this, whilst reference values for the tests examined in the current study exist in many clinical and ageing populations where their use is potentially more suited, circumstances occur where these tests may be used outside of such populations meaning values such as those found in the current study remain important. The diagnosis of certain clinical conditions (e.g. some cancers) may occur across a wide age range whilst not always being accompanied by the presence of other co-morbidities or physiological limitations that some other clinical populations may experience. It is therefore necessary to have reference values to support the pre-existing literature and future studies relating to these age ranges.

In conclusion, this study has demonstrated the test-retest reliability for the TUG, FTSTS, SCT and 6MWT exceeds the ICC threshold of above 0.90 that is required for a clinical test [16] when performed within a 4 week period between sessions in apparently healthy adults aged 30-74 years. Despite research already existing to the test-retest reliability of these tests, there is still limited data regarding measures of absolute reliability, especially when performed with weeks rather than days in between testing sessions. Although not directly related to a specific clinical population, the presentation of measures of absolute reliability such as LOA, SEM and MDC_{95} in the current study adds valuable information to the existing literature. By providing further reference thresholds of absolute reliability, clinicians and researchers alike can use the information to identify meaningful changes beyond those due to measurement error and individual variability. This will aid in assessing the effectiveness of exercise interventions and rehabilitation programmes in settings where more sophisticated facilities and techniques may not be available.

ACKNOWLEDGEMENTS

The authors wish to thank all the volunteers who participated in the study.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

REFERENCES

[1] Samson MM, Meeuwsen IB, Crowe A, Dessens JA, Duursma SA, Verhaar HJ. Relationships between physical performance measures, age, height and body weight in healthy adults. Age Ageing. 2000; 29(3): 235-42.

[2] Hakola L, Komulainen P, Hassinen M, Savonen K, Litmanen H, Lakka TA, et al. Cardiorespiratory fitness in aging men and women: the DR's EXTRA study. Scand J Med Sci Sports. 2011; 21(5): 679-87. doi: 10.1111/j.1600-0838.2010.01127.x

[3] Yümin ET, Şimşek TT, Sertel M, Öztürk A, Yümin M. The effect of functional mobility and balance on health-related quality of life (HRQoL) among elderly people living at home and those living in nursing home. Arch Gerontol Geriatr. 2011; 52(3): e180-4. doi: http://dx.doi.org/10.1016/j.archger.2010.10.027

[4] Carli F. Physiologic considerations of Enhanced Recovery After Surgery (ERAS) programs: implications of the stress response. Can J Anaesth. 2015; 62(2): 110-9. doi: 10.1007/s12630-014-0264-0

[5] Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test. Phys Ther. 2000; 80(9): 896-903.

[6] Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc. 1991; 39(2): 142-8.

[7] Ingle L, Cleland JG, Clark AL. The relation between repeated 6-minute walk test performance and outcome in patients with chronic heart failure. Ann Phys Rehabil Med. 2014; 57(4): 244-53. doi: 10.1016/j.rehab.2014.03.004

[8] Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998; 26(4): 217-38.

[9] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther. 1997; 77(7): 745-50.

[10] Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. Phys Ther. 2006; 86(5): 735-43.

[11] Goldberg A, Chavis M, Watkins J, Wilson T. The five-times-sit-to-stand test: validity, reliability and detectable change in older females. Aging Clin Exp Res. 2012; 24(4): 339-44.

[12] Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports. 2001; 11(5): 280-6.

[13] American Thoracic Society. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med. 2002; 166(1): 111-7. doi: 10.1164/ajrccm.166.1.at1102

[14] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86(2): 420-8.

[15] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476): 307-10.

[16] Portney L, Watkins M. Foundations of clinical research: applications to practice. Prentice Hall, Upper Saddle River. 2008.

[17] Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Amorim-Costa C. The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. J Clin Epidemiol. 2011; 64(3): 264-9. doi: 10.1016/j.jclinepi.2009.11.010

[18] Carli F, Charlebois P, Stein B, Feldman L, Zavorsky G, Kim DJ, et al. Randomized clinical trial of prehabilitation in colorectal surgery. Br J Surg. 2010; 97(8): 1187-97. doi: 10.1002/bjs.7102

[19] Ross RM, Murthy JN, Wollak ID, Jackson AS. The six minute walk test accurately estimates mean peak oxygen uptake. BMC Pulm Med. 2010; 10:31. doi: 10.1186/1471-2466-10-31

[20] Demers C, McKelvie RS, Negassa A, Yusuf S. Reliability, validity, and responsiveness of the six-minute walk test in patients with heart failure. Am Heart J. 2001; 142: 698-703.

[21] Hamilton DM, Haennel RG. Validity and reliability of the 6-minute walk test in a cardiac population. J Cardiopulm Rehabil. 2000; 20(3): 156-64.

[22] Hanson LC, McBurney H, Taylor NF. The retest reliability of the six-minute walk test in patients referred to a cardiac rehabilitation programme. Physiother Res Int. 2012; 17(1): 55-61. doi: 10.1002/pri.513

[23] Alfonso-Rosa RM, Del Pozo-Cruz B, Del Pozo-Cruz J, Sanudo B, Rogers ME. Test-retest reliability and minimal detectable change scores for fitness assessment in older adults with type 2 diabetes. Rehabil Nurs. 2014; 39(5): 260-8. doi: 10.1002/rnj.111.

[24] Harada ND, Chiu V, Stewart AL. Mobility-related function in older adults: assessment with a 6-minute walk test. Arch Phys Med Rehabil. 1999; 80(7): 837-41.

[25] King MB, Judge JO, Whipple R. Wolfson L. Reliability and responsiveness of two physical performance measures examined in the context of functional training intervention. Phys Ther. 2000; 80(1): 8-16.

[26] Kervio G, Carre F, Ville NS. Reliability and intensity of the six-minute walk test in healthy elderly subjects. Med Sci Sports Exerc. 2003; 35(1): 169-74. doi: 10.1249/01.mss.0000043545.02712.a7

[27] Troosters T, Gosselink R, Decramer M. Six minute walking distance in healthy elderly subjects. Eur Respir J. 1999; 14(2): 270-4.

[28] Kristjánsdóttir Á, Ragnarsdóttir M, Einarsson MB, Torfason B. A Comparison of the 6-Minute Walk Test and Symptom Limited Graded Exercise Test for Phase II Cardiac Rehabilitation of Older Adults. J Geriatric Physical Therapy. 2004; 27(2): 65-8.

[29] Montgomery PS, Gardner AW. The clinical utility of a six-minute walk test in peripheral arterial occlusive disease patients. J Am Geriatr Soc. 1998; 46(6): 706-11.

[30] Bohannon RW, Bubela DJ, Magasi SR, Gershon RC. Relative reliability of three objective tests of limb muscle strength. Isokinet Exerc Sci. 2011; 19(2): 77-81. doi: 10.3233/IES-2011-0400

[31] Schaubert K, Bohannon RW. Reliability of the sit-to-stand test over dispersed test sessions. Isokinet Exerc Sci. 2005; 13(2): 119-22.

[32] Schaubert KL, Bohannon RW. Reliability and validity of three strength measures obtained from community-dwelling elderly persons. J Strength Cond Res. 2005; 19(3): 717-20. doi: 10.1519/R-15954.1

[33] Jette AM, Jette DU, Ng J, Plotkin DJ, Bach MA. Are performance-based measures sufficiently reliable for use in multicenter trials? Musculoskeletal Impairment (MSI) Study Group. J Gerontol A Biol Sci Med Sci. 1999; 54(1): M3-6.

[34] Bohannon RW. Test-retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults. J Strength Cond Res. 201; 25(11): 3205-7. doi: 10.1519/JSC.0b013e318234e59f

[35] Rejeski WJ, Ettinger WH, Jr., Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage. 1995; 3(3):157-67.

Table 1. Between session performance differences for the Timed up and go (TUG), Five times sit to stand (FTSTS) and Stair climb test (SCT).

		Session 1 (SD) [Range]	**Session 2 (SD) [Range]**	**Mean difference (SD) [95% CI]**	***P* value**
TUG (s)	Males (n=18)	5.98 (1.41)	5.70 (1.20)	-0.28 (0.38)	0.007
		[4.20 - 8.89]	[4.01 - 8.60]	[-0.46; -0.09]	
	Females (n=17)	6.46 (1.44)	6.31 (1.78)	-0.15 (0.41)	0.159
		[4.21 - 9.21]	[4.12 - 8.41]	[-0.36; 0.06]	
	Combined (n=35)	**6.21 (1.42)**	**6.00 (1.21)**	**-0.21**	**0.003**
		[4.20 - 9.21]	**[4.01 - 8.60]**	**[-0.35; -0.08]**	
FTSTS (s)	Males (n=18)	10.96 (2.86)	10.61 (2.94)	-0.36 (0.38)	0.073
		[6.20 - 17.50]	[5.76 - 17.87]	[-0.75; 0.04]	
	Females (n=17)	11.87 (2.94)	11.33 (2.67)	-0.54 (1.33)	0.113
		[6.45 - 19.64]	[7.07 - 17.74]	[-1.22; 0.14]	
	Combined (n=35)	**11.40 (2.89)**	**10.96 (2.79)**	**-0.44**	**0.019**
		[6.20 - 19.64]	**[9.27 - 17.87]**	**[-0.81; -0.08]**	
SCT (s)	Males (n=18)	2.79 (0.45)	2.73 (0.46)	-0.05 (0.11)	0.048
		[2.13 - 3.68]	[2.03 - 3.61]	[-0.11; -0.00]	
	Females (n=17)	2.85 (0.51)	2.80 (0.58)	-0.04 (0.19)	0.348
		[1.93 - 3.69]	[1.71 – 3.83]	[-0.14; 0.05]	
	Combined (n=35)	**2.82 (0.48)**	**2.77 (0.51)**	**-0.05**	**0.061**
		[1.93 - 3.69]	**[1.71 - 3.83]**	**[-0.10; +0.01**	

SD: standard deviation; 95% CI: 95% confidence intervals; s: seconds

Table 2. Reliability data for the Timed up and go (TUG), Five times sit to stand (FTSTS) and Stair climb test (SCT).

		$ICC_{3,2}$ [95% CI]	95% LOA	SEM	SEM%	MDC_{95}	$MDC_{95\%}$
TUG (s)	Males (n=18)	0.97 (0.86 - 0.99)	-1.02; +0.47	0.23	3.89	0.63	10.79
	Females (n=17)	0.97 (0.92 - 0.99)	-0.95; +0.63	0.22	3.69	0.60	9.33
	Combined (n=35)	**0.97 [0.93 - 0.99]**	-0.99; +0.56	0.22	3.67	0.62	10.18
FTSTS (s)	Males (n=18)	0.98 (0.94 - 0.99)	-1.90; +1.19	0.43	3.94	1.18	10.92
	Females (n=17)	0.94 (0.82 - 0.98)	-3.14; +2.06	0.71	6.12	1.96	16.92
	Combined (n=35)	**0.96 [0.91 - 0.98]**	-2.55; +1.66	0.58	5.19	1.60	16.09
SCT (s)	Males (n=18)	0.98 (0.95 - 0.99)	-0.27; +0.16	0.06	2.13	0.16	5.91
	Females (n=17)	0.97 (0.92 - 0.99)	-0.41; +0.33	0.09	3.32	0.26	9.21
	Combined (n=35)	**0.98 [0.95 - 0.99]**	-0.34; +0.25	0.08	2.80	0.22	7.77

ICC: Intraclass correlation; 95% CI: 95% confidence interval; 95% LOA: 95% limit of agreements; SEM: Standard error of measurement; MDC_{95}: Minimum detectable change at the 95% confidence interval

Table 3. Between session performance and physiological differences for the 6 minute walk test (6MWT).

6 Minute Walk Test (6MWT)					
		Session 1 (SD) [Range]	**Session 2 (SD) [Range]**	**Mean difference [95% CI]**	***P* value**
Distance walked (m)	Males (n=18)	613.2 (73.9) [486 - 726]	625.3 (86.9) [483 - 759]	+12.1 (20.7) [1.8; 22.4]	0.024
	Females (n=17)	576.7 (78.3) [437 - 699]	575.5 (75.1) [451 - 705]	-1.2 (14.5) [-8.7; 6.2]	0.729
	Combined (n=35)	**595.5 (77.2) [437 - 726]**	**601.1 (84.1) [451 - 759]**	**+5.6 (18.9) [-0.87; +12.13]**	**0.087**
HR_{pre} (bpm)	Males (n=18)	68.1 (12.2) [52 - 94]	69.7 (9.5) [54 - 84]	1.6 (9.6) [-3.2; 6.4]	0.503
	Females (n=17)	72.3 (13.4) [52 - 98]	68.6 (10) [52 – 88]	-3.7 (9.3) [-8.5; 1.1]	0.119
	Combined (n=35)	**70.1 (12.8) [52 - 98]**	**69.1 (9.6) [52 - 88]**	**-1.0 (9.7) [-4.33; +2.33]**	**0.546**
HR_{post} (bpm)	Males (n=18)	107.4 (22.0) [78 - 165]	110.6 (23.4) [71 – 161]	13.2 (11.2) [-2.4; 8.6]	0.248
	Females (n=17)	112.4 (24.5) [76 - 166]	110.7 (23.7) [80 – 159]	-1.7 (6.2) [-4.9; 1.5]	0.275
	Combined (n=35)	**109.8 (23.1) [83.5 - 157.0]**	**110.6 (23.2) [71.0 - 161.0]**	**+0.8 (9.3) [-2.4; +4.0]**	**0.616**
HR_{ave} (bpm)	Males (n=18)	109.1 (20.1) [83.5 - 157.0]	110.3 (21.2) [75 - 151]	1.2 (7.9) [-2.7; 5.2]	0.522
	Females (n=17)	112.6 (16.8) [84 - 140]	111.4 (17.2) [84 - 142]	-1.3 (6.4) [-4.6; 2.0]	0.420
	Combined (n=35)	**110.8 (18.4) [83.5 - 157.0]**	**110.8 (19.1) [75.0 - 151.0]**	**+0.0 (7.2) [-2.5; +2.5]**	**0.998**
$SaO2_{pre}$ (%)	Males (n=18)	97.9 (1.0) [96 - 99]	96.9 (1.6) [94 - 99]	-1.0 (2.0) [-2.0; - 0.0]	0.046
	Females (n=17)	98.0 (1.1) [95 - 100]	97.5 (1.6) [94 - 100]	-0.5 (1.3) [-1.2; 0.3]	0.187
	Combined (n=35)	**97.9 (1.0) [95 - 100]**	**97.2 (1.7) [94 - 100]**	**-0.5 (1 [-1.0; +0.5]**	**0.073**
$SaO2_{post}$ (%)	Males (n=18)	97.7 (1.5) [93 - 100]	96.7 (1.8) [91 - 98]	-0.9 (2.3) [-2.1; 0.2]	0.094
	Females (n=17)	97.0 (2.6) [89 - 99]	97.5 (2.2) [91 - 100]	0.5 (1.3) [-0.2; 1.3]	0.135
	Combined (n=35)	**97.4 (2.1) [89 - 100]**	**97.2 (2.0) [91 - 100]**	**-0.2 [-0.9; +0.5]**	**0.552**

HR_{pre}: Heart rate prior to 6MWT; HR_{post}: Heart rate post 6MWT; HR_{ave}: Average heart rate; $SaO2_{pre}$: Oxygen saturation prior to 6MWT; $SaO2_{post}$: Oxygen saturation post 6MWT

Table 4. Reliability data for the 6 minute walk test (6MWT)

		$ICC_{3,2}$ [95% CI]	95% LOA	SEM	SEM%	MDC_{95}	$MDC_{95\%}$
6MWT (m)	Males (n=18)	0.96 (0.86 - 0.99)	-28.4; +52.6	16.3	2.6	45.3	7.3
	Females (n=17)	0.98 (0.95 - 0.99)	-29.6; +27.1	9.9	1.7	27.3	4.7
	Combined (n=35)	**0.97 [0.94 - 0.99]**	**-31.4; +42.7**	**13.7**	**2.3**	**37.8**	**6.3**

ICC: Intraclass correlation; 95% CI: 95% confidence interval; 95% LOA: 95% limit of agreements; SEM: Standard error of measurement; MDC_{95}: Minimum detectable change at the 95% confidence interval