

This is a repository copy of *The test–retest reliability of four functional mobility tests in apparently healthy adults*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/100376/>

Version: Accepted Version

---

**Article:**

Northgraves, Matthew [orcid.org/0000-0001-9260-8643](https://orcid.org/0000-0001-9260-8643), Hayes, Stephen, Marshall, Philip et al. (2 more authors) (2016) The test–retest reliability of four functional mobility tests in apparently healthy adults. *Isokinetics and Exercise Science*. pp. 171-179. ISSN 1878-5913

<https://doi.org/10.3233/IES-160614>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

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

3

4 **Authors:** Matthew J. Northgraves<sup>1,2\*</sup>, Stephen C. Hayes<sup>1</sup>, Philip Marshall<sup>1</sup>, Leigh A.  
5 Madden<sup>3</sup>, and Rebecca V. Vince<sup>1</sup>

6 <sup>1</sup>Department of Sport, Health and Exercise Science, University of Hull, Hull, UK

7 <sup>2</sup>Department of Health Sciences, University of York, York, UK

8 <sup>3</sup>School of Biological, Biomedical and Environmental Sciences, University of Hull,  
9 Hull, UK

10

11 **\*Correspondence:**

12 Mr Matthew J. Northgraves

13 Department of Health Sciences, University of York, York, UK, YO10 5DD.

14 Telephone: +44 (0)1904 321614

Email: [matt.northgraves@york.ac.uk](mailto:matt.northgraves@york.ac.uk)

15

16

17

18

19

20 **ABSTRACT**

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

24 **OBJECTIVES:** To examine the test-retest and absolute reliability of timed up and go  
25 test (TUG), five times sit-to-stand (FTSTS), stair climb test (SCT) and 6 minute walk  
26 (6MWT).

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

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

35 **CONCLUSIONS:** The results of this study indicate that the TUG, FTSTS, SCT and  
36 6MWT are reliable when performed four weeks apart. Furthermore, the inclusion of  
37 SEM, MDC and 95% LOA provides reference values to aid in identifying changes over  
38 time above those of measurement error and individual variation.

39

40 **KEYWORDS:** TUG, FTSTS, Stair climb, 6MWT

41 **1. INTRODUCTION**

42

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

51

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

63 to change over longer periods is potentially compromised by the presence of  
64 measurement error and variation in individual performance.

65

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

75

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

79

## 80 **2. METHODS**

### 81 **2.1. SUBJECTS**

82 A sample of 35 volunteers (18 males, 17 females) was recruited from the local  
83 community via advertisement for this study. Inclusion criteria included being an

84 apparently healthy male and female aged 30-75 years. Exclusion criteria included any  
85 history of cardiopulmonary conditions, any musculoskeletal and/or orthopaedic  
86 conditions, current injury, history of fracture within the last year, uncorrected visual  
87 impairment, recent history of dizziness or fainting, vestibular disorders and shortness of  
88 breath with minimum exertion. Participants were screened for eligibility through the  
89 completion of an institution approved pre-exercise medical questionnaire. All  
90 participants provided written informed consent, and the study was approved by the  
91 Department of Sport, Health and Exercise Human Ethics Committee and followed the  
92 principles outlined in the Declaration of Helsinki.

93

## 94 2.2. EXPERIMENTAL DESIGN

95 As the purpose of this study was to test the test-retest reliability of the four assessment  
96 measures rather than inter-rater reliability, all trials were conducted by a single tester;  
97 this ensured maximum consistency for data collection of each variable. Participants  
98 were required to attend two identical testing sessions separated by approximately four  
99 weeks. Both sessions were conducted at the same time of day in order to control for  
100 circadian variation and participants were asked to refrain from strenuous exercise in the  
101 24 hours preceding each visit. The order of testing was the TUG, followed by the five  
102 times sit to stand (FTSTS), stair climb test (SCT) and finally the 6MWT.

103

104 1. TUG: From a plastic chair measuring 40 cm from the floor and 39 cm deep,  
105 participants were asked to stand from a seated position, walk 3 metres before turning  
106 180° and returning to the chair to sit down. Timing started with the count of “THREE,

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

110 2. FTSTS: Using a chair as above, participants were instructed the aim of the test was to  
111 perform five sit to stand movements as fast as they could in a controlled safe manner.  
112 From an upright seated position with their back against the chair backrest and arms  
113 crossed over their chest, the test started with the count of “THREE, TWO, ONE, GO”  
114 [11].

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

121 4. 6MWT: A 30 metre flat walking surface was set out with cones marking each 3  
122 metre interval with distinct markers at the start and end. Following a period of 10  
123 minutes seated rest, participants were instructed to walk as far and as fast as possible in  
124 6 minutes. Rest periods were permitted however time was not stopped. A standardised  
125 protocol was used in line with the guidelines provided by the ATS [13]. At the end of  
126 the 6 minutes, participants stopped when instructed with the total distance walked  
127 providing the primary outcome measure. Measures of heart rate (HR) and arterial  
128 oxygen saturation (SaO<sub>2</sub>) (Nonin Onyx finger pulse Oximeter, Nonin Medical Inc,  
129 Plymouth, Minnesota) were taken prior to (HR<sub>pre</sub>, SaO<sub>2pre</sub>) and immediately after the

130 6MWT ( $HR_{\text{post}}$ ,  $SaO_{2\text{post}}$ ). Heart rate was measured at one minute intervals throughout  
131 the test allowing the average HR ( $HR_{\text{ave}}$ ) to be calculated.

132

133 For TUG, FTSTS, and SCT, following an unrecorded familiarisation trial, the mean of  
134 three trials were taken for analysis. A single trial per session was performed for the  
135 6MWT.

136

### 137 2.3. STATISTICAL ANALYSIS

138

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

145

146 Relative reliability was assessed using the ICC model 3 [14]. As the mean of three trials  
147 was used for the TUG, FTSTS and SCT, test-retest reliability was measured using  
148  $ICC_{3,2}$  model. For the 6MWT, which involved a single trial each session the  $ICC_{3,1}$   
149 model was used. Absolute reliability was expressed using 95% limits of agreement  
150 (95% LOA) [15], SEM and minimum detectable change at a 95% confidence interval



151 (MDC<sub>95</sub>). The 95% LOA represents the expected range of difference scores for each  
152 test. The SEM allowed measurement error to be displayed in the same units as the  
153 original measurement and was calculated using the formula:

$$154 \quad \text{SEM} = \text{SD} \times \sqrt{(1-\text{ICC})}$$

155 where SD was the standard deviation for all observations from test sessions 1 and 2 and  
156 ICC was the reliability coefficient. Measurement error was also expressed as a  
157 percentage of the mean (SEM<sub>%</sub>) using the formula:

$$158 \quad \text{SEM}_{\%} = (\text{SEM}/\text{mean}) \times 100$$

159 This represents the smallest change required to indicate real change in a group of  
160 participants. MDC<sub>95</sub> was calculated to represent the magnitude of change required to  
161 exceed the anticipated measurement variation, measurement error and variability of  
162 participants with 95% confidence [10]. The formula used for calculating MDC<sub>95</sub> was:

$$163 \quad \text{MDC}_{95} = \text{SEM} \times 1.96 \times \sqrt{2}$$

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

167

### 168 **3. RESULTS**

169

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

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

177

### 178 3.1. TUG, FTSTS AND SCT

179

180 A mean percentage improvement in the performance time of TUG (3.4%; Range: -10.4  
181 to +16.0%), FTSTS (3.9%; Range: +20.5 to -23.7%) and SCT (1.7%; Range: +12.4 to -  
182 0.3%) was seen between the first and second visit. The improvement however was only  
183 significant ( $p < 0.05$ ) for the TUG and FTSTS (Table 1). The results relating to the both  
184 relative (ICC) and absolute reliability (LOA, SEM & MDC) of the TUG, FTSTS and  
185 SCT are displayed in Table 2. All three tests demonstrated good test-retest reliability  
186 with high ICCs ranging from 0.96 to 0.98. Out of the three tests, the SCT displayed the  
187 greatest absolute reliability with the SEM represented as a percentage of the mean being  
188 2.8% whilst the FTSTS had the greatest measurement error at 5.8% of the mean.

189

190 When analysed based on gender, mean performance time for all three tests was faster in  
191 males (Table 1), however neither relative nor absolute reliability were greatly affected  
192 (Table 2). The magnitude of the ICCs for all three tests remained similar in males (ICCs  
193 = 0.97 to 0.98) and females (ICCs = 0.94 to 0.97) compared to when all participants

194 were combined (ICCs 0.96 to 0.98). In respects to absolute reliability, the greatest  
195 variability between genders was observed in the FTSTS.

196

### 197 3.2. 6MWT

198

199 A mean improvement of approximately 5.6 metres (+0.9%) was seen between the first  
200 and second visit although this was not significant ( $p > 0.05$ ) (Table 3). No significant  
201 difference was seen between sessions for  $\text{SaO}_{2\text{post}}$ ,  $\text{HR}_{\text{pre}}$ ,  $\text{HR}_{\text{post}}$  or  $\text{HR}_{\text{ave}}$  however  
202  $\text{SaO}_{2\text{pre}}$  was significantly lower in session 2. The high ICC and narrow accompanying  
203 95% CI demonstrated good test-retest reliability for the 6MWT (Table 4). Furthermore,  
204 the values reported for both 95% LOA and  $\text{MDC}_{95}$  were similar whilst the SEM of 13.7  
205 metres ( $\text{SEM}_{\%} -2.3\%$ ) represented a low value of measurement error.

206

207 When analysed based on gender, the mean distance walked was significantly further  
208 (+12.1 metres; +2.0%;  $p < 0.05$ ) in the 2<sup>nd</sup> session for males however no difference  
209 between sessions was evident for females (-1.2 metres; 0.2%;  $p > 0.05$ ). Despite the  
210 difference in males between sessions neither the relative nor absolute reliability of the  
211 6MWT was greatly affected.

212

## 213 4. DISCUSSION

214

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

227

228 Of the four tests included in the current study, the 6MWT is probably the most  
229 frequently used acting as a means of assessing the effectiveness of different intervention  
230 programmes [18] as well as a predictor of both cardiorespiratory fitness [19] and  
231 clinical outcomes [7]. As in the current study, good test-retest reliability has been  
232 observed in a number of other populations including cardiac patients (ICCs = 0.88 -  
233 0.97) [20-22], type 2 diabetics (ICC = 0.99) [23] and the elderly (ICCs = 0.87 – 0.93)  
234 [24]. It is however often reported that at least one, if not more, familiarisation trials are  
235 required in order to alleviate any potential learning effect and thus achieve a consistent  
236 baseline measurement for the 6MWT [21, 22, 26, 27].

237

238 In healthy individuals aged 60-70; it was only from the third trial that the measurement  
239 became reliable when performing five 6MWT over a 1 week period [26]. Between both  
240 the 1st and 2nd, and 2nd and 3rd trials a mean increase of ~20 metres was reported;  
241 representing a 3.7 – 3.8% increase between trials. An average improvement of  $8 \pm 5\%$   
242 (+47 metres) in the second of two trials performed on the same day was observed in  
243 healthy individuals aged 50 - 85 years [27]. Both Hanson et al. [22] and Hamilton et al.  
244 [21] reported a learning effect occurred between trials within a cardiac rehabilitation  
245 setting despite reporting good relative reliability (ICC=0.91 and 0.97 respectively). An  
246 11.8% (+52 metres) increase in distance walked was observed in Hanson et al. [22]  
247 between the 1<sup>st</sup> and 2<sup>nd</sup> trial and this increased to 19.1% (+85 metres) between the 1<sup>st</sup>  
248 and 3<sup>rd</sup> trial. Furthermore, whether the three tests were performed on the same day or  
249 spread over a week did not alter the presence of the learning effect [22]. Although the  
250 improvement was smaller, Hamilton et al. [21] observed a 3.5% (+18 metres) increase  
251 between the 1<sup>st</sup> and 2<sup>nd</sup> trial and 5.6% (+29 metres) between the 1<sup>st</sup> and 3<sup>rd</sup> trial.

252

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

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

266

267 It is acknowledged that direct comparisons between this study and those using clinical  
268 populations are difficult as considerable variation does exist between population groups.  
269 The SEM (13.7 metres) and  $MDC_{95}$  (37.8 metres) seen in the current study were  
270 comparable to those reported in older type 2 diabetics (SEM = 9.88 metres;  $MDC_{95}$  =  
271 27.37 metres) by Alfonsa-Rosa et al. [23]. This was despite only a 1 week period  
272 existing between their trials suggesting any learning effect was absent in their study  
273 [23]. These values however do differ from those seen in both elderly (SEM: 32-34  
274 metres;  $MDC_{95}$ : 88.7-95 metres [24] and cardiac (SEM: 18.4-32.6 metres;  $MDC_{95}$ :  
275 50.92 – 90.3 [20, 21, 29] populations therefore patient characteristics and conditions  
276 need to be considered in determining changes in performance.

277

278 Unlike with the 6MWT, the presence of a significant statistical decrease in time taken to  
279 perform the TUG and FTSTS between the first and second sessions suggested a learning  
280 effect was present. Similar improved FTSTS performance times have previously been  
281 reported in trials separated by 4-10 days [30] up to six weeks [31, 32]. Despite this, the  
282 ICC for all three studies was in excess of 0.80 indicating good correlation and  
283 agreement between trials. The ICC of 0.97 for the TUG in the current study (Table 1)  
284 exceeded that of Jette et al. [33], who reported an ICC of 0.74 in elderly frail

285 individuals. However, the difference in study populations is likely to have influenced  
286 the reduced ICC in Jette et al. [33] compared to the current study. It is also worth noting  
287 that whilst the median number of days between trials was 14 days in Jette et al. [33], the  
288 overall range between trials varied from 0 days to 132 days. It is therefore plausible that  
289 the decrease in test-retest reliability, as indicated by ICC, was related to a true change in  
290 the study populations' ability to perform the FTSTS; especially in the individuals with  
291 the largest number of days between trials.

292

293 The results relating to the relative reliability of the FTSTS when performed with an  
294 extended period between trials have previously been varied [34]. In trials separated by  
295 4-10 days, Bohannon et al [30] reported good test-retest reliability (ICC = 0.96; 95%  
296 CI: 0.92-0.98) in community-dwelling men and women aged 15-85 years. In contrast,  
297 when the interval between trials has been longer, lower ICC's have tended to be  
298 reported. In two studies by Schaubert and Bohannon [31, 32] in which testing sessions  
299 were separated by 6 weeks, ICCs of 0.82 (95% CI: 0.68-0.92) and 0.81 (95% CI: not  
300 stated) respectively were reported. In the current study, despite the 4 week period  
301 between tests, test-retest reliability remained good with the ICC of 0.96 far exceeding  
302 those seen in the two aforementioned studies.

303

304 This difference could potentially be explained by a number of factors, including the  
305 presence of a shorter four week period between testing sessions in the current study as  
306 opposed to six weeks [31, 32]. Furthermore, the sample sizes used in both these studies  
307 (n=21 [31] and n=11 [32]) were smaller than those of the current study (n=35). A more

308 pertinent factor however is probably the difference in participant ages between the  
309 studies. It is acknowledged that the mean ages in both Schaubert and Bohannon studies  
310 [31, 32] ( $75.0 \pm 5.9$  years [Range: 65-85 years] and  $75.5 \pm 5.8$  years [Range: 65-85  
311 years] respectively) make their findings more generalizable, especially to older  
312 populations where the FTSTS is more traditionally used, than the current study ( $54.6 \pm$   
313  $12.1$  years [Range: 30 -74 years]). Despite this, the current study adds to the existing  
314 literature with regards to the potential measurement error of the four tests investigated.

315

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

327

328 The use of 95% LOA as a measure of absolute reliability in the current study reflected  
329 the most conservative method. For the FTSTS, 95% LOA suggested a change of over  
330 2.55 seconds was required to detect real change compared to the 1.60 seconds according



331 to the  $MDC_{95}$  (Table 2). Understanding the variation present in both the performance of  
332 the test and the different methods of calculating absolute reliability could be important  
333 when assessing any change present in repeated performances.

334

335 Although in the current study the SCT displayed good relative test-retest reliability  
336 ( $ICC = 0.98$ ; 95% CI 0.95-0.99) and absolute reliability ( $SEM = 0.08$  s;  $MDC_{95} = 0.22$   
337 s), the results remain difficult to interpret. Variations of the SCT have been used in a  
338 variety of different populations including those with orthopaedic limitations and the  
339 elderly. The intra-session reliability in elderly individuals (mean age 69.4 years) with  
340 hip and/or knee osteoarthritis was reported to be good with an  $ICC$  of 0.94 (95% CI  
341 0.75-0.98) and  $SEM$  of 0.28 seconds seen for a four step ascent only SCT [12]. When  
342 performing a five step SCT including both the ascent and descent of the stairs two  
343 weeks apart, Rejeski et al. [35] reported good test- retest reliability ( $ICC = 0.93$ ; 95% CI  
344 Not reported) in patients with knee osteoarthritis. Despite similar  $ICC$  being reported in  
345 Lin et al. [12], Rejeski et al. [35] and the current study, making comparisons between  
346 the studies is difficult. The absence of any limiting condition such as osteoarthritis in  
347 the present study that may have impaired the ability of participants to climb stairs,  
348 means the performance time of 2.77 seconds is faster than those reported in either Lin et  
349 al. [12] ( $4.17 \pm 2.80$  s) or Rejeski et al. [35] ( $10.21 \pm 4.45$  s). It is therefore  
350 acknowledged the SCT results are difficult to generalise beyond the present study.

351

352 This study is not without limitations. The use of an apparently healthy population with a  
353 relatively wide age range (30-74 years) in this study means the results cannot be directly

354 generalised to those of a specific clinical population. Furthermore, given the sample  
355 size, stratification based on factors such as age, gender and self-reported physical  
356 activity was not possible. The sub-analysis based on gender alone (Tables 2 and 4) did  
357 not differ greatly between the genders for any of the tests in the current study, however  
358 whether a more pronounced difference would be observed with a larger sample size  
359 cannot be dismissed.

360

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

370

371 In conclusion, this study has demonstrated the test-retest reliability for the TUG,  
372 FTSTS, SCT and 6MWT exceeds the ICC threshold of above 0.90 that is required for a  
373 clinical test [16] when performed within a 4 week period between sessions in apparently  
374 healthy adults aged 30-74 years. Despite research already existing to the test-retest  
375 reliability of these tests, there is still limited data regarding measures of absolute  
376 reliability, especially when performed with weeks rather than days in between testing

377 sessions. Although not directly related to a specific clinical population, the presentation  
378 of measures of absolute reliability such as LOA, SEM and MDC<sub>95</sub> in the current study  
379 adds valuable information to the existing literature. By providing further reference  
380 thresholds of absolute reliability, clinicians and researchers alike can use the  
381 information to identify meaningful changes beyond those due to measurement error and  
382 individual variability. This will aid in assessing the effectiveness of exercise  
383 interventions and rehabilitation programmes in settings where more sophisticated  
384 facilities and techniques may not be available.

385

#### 386 ACKNOWLEDGEMENTS

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

388

#### 389 CONFLICT OF INTEREST

390 The authors declared no conflict of interest.

391

#### 392 REFERENCES

393 [1] Samson MM, Meeuwse IB, Crowe A, Dessens JA, Duursma SA, Verhaar HJ.  
394 Relationships between physical performance measures, age, height and body weight in healthy  
395 adults. *Age Ageing*. 2000; 29(3): 235-42.

- 396 [2] Hakola L, Komulainen P, Hassinen M, Savonen K, Litmanen H, Lakka TA, et al.  
397 Cardiorespiratory fitness in aging men and women: the DR's EXTRA study. *Scand J Med Sci*  
398 *Sports*. 2011; 21(5): 679-87. doi: 10.1111/j.1600-0838.2010.01127.x
- 399 [3] Yümin ET, Şimşek TT, Sertel M, Öztürk A, Yümin M. The effect of functional mobility and  
400 balance on health-related quality of life (HRQoL) among elderly people living at home and  
401 those living in nursing home. *Arch Gerontol Geriatr*. 2011; 52(3): e180-4. doi:  
402 <http://dx.doi.org/10.1016/j.archger.2010.10.027>
- 403 [4] Carli F. Physiologic considerations of Enhanced Recovery After Surgery (ERAS) programs:  
404 implications of the stress response. *Can J Anaesth*. 2015; 62(2): 110-9. doi: 10.1007/s12630-  
405 014-0264-0
- 406 [5] Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in  
407 community-dwelling older adults using the Timed Up & Go Test. *Phys Ther*. 2000; 80(9): 896-  
408 903.
- 409 [6] Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for  
410 frail elderly persons. *J Am Geriatr Soc*. 1991; 39(2): 142-8.
- 411 [7] Ingle L, Cleland JG, Clark AL. The relation between repeated 6-minute walk test  
412 performance and outcome in patients with chronic heart failure. *Ann Phys Rehabil Med*. 2014;  
413 57(4): 244-53. doi: 10.1016/j.rehab.2014.03.004
- 414 [8] Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in  
415 variables relevant to sports medicine. *Sports Med*. 1998; 26(4): 217-38.
- 416 [9] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an  
417 applied example using elbow flexor strength data. *Phys Ther*. 1997; 77(7): 745-50.

- 418 [10] Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in  
419 physical therapy. *Phys Ther.* 2006; 86(5): 735-43.
- 420 [11] Goldberg A, Chavis M, Watkins J, Wilson T. The five-times-sit-to-stand test: validity,  
421 reliability and detectable change in older females. *Aging Clin Exp Res.* 2012; 24(4): 339-44.
- 422 [12] Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and  
423 hip osteoarthritis. *Scand J Med Sci Sports.* 2001; 11(5): 280-6.
- 424 [13] American Thoracic Society. ATS statement: guidelines for the six-minute walk test. *Am J*  
425 *Respir Crit Care Med.* 2002; 166(1): 111-7. doi: 10.1164/ajrccm.166.1.at1102
- 426 [14] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol*  
427 *Bull.* 1979; 86(2): 420-8.
- 428 [15] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods  
429 of clinical measurement. *Lancet.* 1986; 1(8476): 307-10.
- 430 [16] Portney L, Watkins M. *Foundations of clinical research: applications to practice.* Prentice  
431 Hall, Upper Saddle River. 2008.
- 432 [17] Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Amorim-Costa C. The limits  
433 of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation  
434 of agreement. *J Clin Epidemiol.* 2011; 64(3): 264-9. doi: 10.1016/j.jclinepi.2009.11.010
- 435 [18] Carli F, Charlebois P, Stein B, Feldman L, Zavorsky G, Kim DJ, et al. Randomized clinical  
436 trial of prehabilitation in colorectal surgery. *Br J Surg.* 2010; 97(8): 1187-97. doi:  
437 10.1002/bjs.7102
- 438 [19] Ross RM, Murthy JN, Wollak ID, Jackson AS. The six minute walk test accurately  
439 estimates mean peak oxygen uptake. *BMC Pulm Med.* 2010; 10:31. doi: 10.1186/1471-2466-  
440 10-31

- 441 [20] Demers C, McKelvie RS, Negassa A, Yusuf S. Reliability, validity, and responsiveness of  
442 the six-minute walk test in patients with heart failure. *Am Heart J.* 2001; 142: 698-703.
- 443 [21] Hamilton DM, Haennel RG. Validity and reliability of the 6-minute walk test in a cardiac  
444 population. *J Cardiopulm Rehabil.* 2000; 20(3): 156-64.
- 445 [22] Hanson LC, McBurney H, Taylor NF. The retest reliability of the six-minute walk test in  
446 patients referred to a cardiac rehabilitation programme. *Physiother Res Int.* 2012; 17(1): 55-61.  
447 doi: 10.1002/pri.513
- 448 [23] Alfonso-Rosa RM, Del Pozo-Cruz B, Del Pozo-Cruz J, Sanudo B, Rogers ME. Test-retest  
449 reliability and minimal detectable change scores for fitness assessment in older adults with type  
450 2 diabetes. *Rehabil Nurs.* 2014; 39(5): 260-8. doi: 10.1002/rmj.111.
- 451 [24] Harada ND, Chiu V, Stewart AL. Mobility-related function in older adults: assessment  
452 with a 6-minute walk test. *Arch Phys Med Rehabil.* 1999; 80(7): 837-41.
- 453 [25] King MB, Judge JO, Whipple R, Wolfson L. Reliability and responsiveness of two  
454 physical performance measures examined in the context of functional training intervention.  
455 *Phys Ther.* 2000; 80(1): 8-16.
- 456 [26] Kervio G, Carre F, Ville NS. Reliability and intensity of the six-minute walk test in healthy  
457 elderly subjects. *Med Sci Sports Exerc.* 2003; 35(1): 169-74. doi:  
458 10.1249/01.mss.0000043545.02712.a7
- 459 [27] Troosters T, Gosselink R, Decramer M. Six minute walking distance in healthy elderly  
460 subjects. *Eur Respir J.* 1999; 14(2): 270-4.
- 461 [28] Kristjánsdóttir Á, Ragnarsdóttir M, Einarsson MB, Torfason B. A Comparison of the 6-  
462 Minute Walk Test and Symptom Limited Graded Exercise Test for Phase II Cardiac  
463 Rehabilitation of Older Adults. *J Geriatric Physical Therapy.* 2004; 27(2): 65-8.

464 [29] Montgomery PS, Gardner AW. The clinical utility of a six-minute walk test in peripheral  
465 arterial occlusive disease patients. *J Am Geriatr Soc.* 1998; 46(6): 706-11.

466 [30] Bohannon RW, Bubela DJ, Magasi SR, Gershon RC. Relative reliability of three objective  
467 tests of limb muscle strength. *Isokinet Exerc Sci.* 2011; 19(2): 77-81. doi: 10.3233/IES-2011-  
468 0400

469 [31] Schaubert K, Bohannon RW. Reliability of the sit-to-stand test over dispersed test sessions.  
470 *Isokinet Exerc Sci.* 2005; 13(2): 119-22.

471 [32] Schaubert KL, Bohannon RW. Reliability and validity of three strength measures obtained  
472 from community-dwelling elderly persons. *J Strength Cond Res.* 2005; 19(3): 717-20. doi:  
473 10.1519/R-15954.1

474 [33] Jette AM, Jette DU, Ng J, Plotkin DJ, Bach MA. Are performance-based measures  
475 sufficiently reliable for use in multicenter trials? Musculoskeletal Impairment (MSI) Study  
476 Group. *J Gerontol A Biol Sci Med Sci.* 1999; 54(1): M3-6.

477 [34] Bohannon RW. Test-retest reliability of the five-repetition sit-to-stand test: a systematic  
478 review of the literature involving adults. *J Strength Cond Res.* 2011; 25(11): 3205-7. doi:  
479 10.1519/JSC.0b013e318234e59f

480 [35] Rejeski WJ, Ettinger WH, Jr., Schumaker S, James P, Burns R, Elam JT. Assessing  
481 performance-related disability in patients with knee osteoarthritis. *Osteoarthritis Cartilage.*  
482 1995; 3(3):157-67.

483

484

485

486

487 Table 1. Between session performance differences for the Timed up and go (TUG), Five  
 488 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) [4.20 - 8.89]	5.70 (1.20) [4.01 - 8.60]	-0.28 (0.38) [-0.46; -0.09]	0.007
	Females (n=17)	6.46 (1.44) [4.21 - 9.21]	6.31 (1.78) [4.12 - 8.41]	-0.15 (0.41) [-0.36; 0.06]	0.159
	<b>Combined (n=35)</b>	<b>6.21 (1.42) [4.20 - 9.21]</b>	<b>6.00 (1.21) [4.01 - 8.60]</b>	<b>-0.21 [-0.35; -0.08]</b>	<b>0.003</b>
FTSTS (s)	Males (n=18)	10.96 (2.86) [6.20 - 17.50]	10.61 (2.94) [5.76 - 17.87]	-0.36 (0.38) [-0.75; 0.04]	0.073
	Females (n=17)	11.87 (2.94) [6.45 - 19.64]	11.33 (2.67) [7.07 - 17.74]	-0.54 (1.33) [-1.22; 0.14]	0.113
	<b>Combined (n=35)</b>	<b>11.40 (2.89) [6.20 - 19.64]</b>	<b>10.96 (2.79) [9.27 - 17.87]</b>	<b>-0.44 [-0.81; -0.08]</b>	<b>0.019</b>
SCT (s)	Males (n=18)	2.79 (0.45) [2.13 - 3.68]	2.73 (0.46) [2.03 - 3.61]	-0.05 (0.11) [-0.11; -0.00]	0.048
	Females (n=17)	2.85 (0.51) [1.93 - 3.69]	2.80 (0.58) [1.71 - 3.83]	-0.04 (0.19) [-0.14; 0.05]	0.348
	<b>Combined (n=35)</b>	<b>2.82 (0.48) [1.93 - 3.69]</b>	<b>2.77 (0.51) [1.71 - 3.83]</b>	<b>-0.05 [-0.10; +0.01]</b>	<b>0.061</b>

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

489

490

491

492

493

494

495



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

		ICC <sub>3,2</sub> [95% CI]	95% LOA	SEM	SEM%	MDC <sub>95</sub>	MDC <sub>95%</sub>
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
	<b>Combined (n=35)</b>	<b>0.97 [0.93 - 0.99]</b>	-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
	<b>Combined (n=35)</b>	<b>0.96 [0.91 - 0.98]</b>	-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
	<b>Combined (n=35)</b>	<b>0.98 [0.95 - 0.99]</b>	-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<sub>95</sub>: Minimum detectable change at the 95% confidence interval

498

499

500

501

502

503

504

505

506

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

		6 Minute Walk Test (6MWT)			
		Session 1	Session 2	Mean difference	P value
		(SD)	(SD)		
		[Range]	[Range]	[95% CI]	
Distance walked (m)	Males (n=18)	613.2 (73.9)	625.3 (86.9)	+12.1 (20.7)	0.024
	Females (n=17)	576.7 (78.3)	575.5 (75.1)	-1.2 (14.5)	0.729
	<b>Combined (n=35)</b>	<b>595.5 (77.2)</b>	<b>601.1 (84.1)</b>	<b>+5.6 (18.9)</b>	<b>0.087</b>
		<b>[437 - 726]</b>	<b>[451 - 759]</b>	<b>[-0.87; +12.13]</b>	
HR <sub>pre</sub> (bpm)	Males (n=18)	68.1 (12.2)	69.7 (9.5)	1.6 (9.6)	0.503
	Females (n=17)	72.3 (13.4)	68.6 (10)	-3.7 (9.3)	0.119
	<b>Combined (n=35)</b>	<b>70.1 (12.8)</b>	<b>69.1 (9.6)</b>	<b>-1.0 (9.7)</b>	<b>0.546</b>
		<b>[52 - 98]</b>	<b>[52 - 88]</b>	<b>[-4.33; +2.33]</b>	
HR <sub>post</sub> (bpm)	Males (n=18)	107.4 (22.0)	110.6 (23.4)	13.2 (11.2)	0.248
	Females (n=17)	112.4 (24.5)	110.7 (23.7)	-1.7 (6.2)	0.275
	<b>Combined (n=35)</b>	<b>109.8 (23.1)</b>	<b>110.6 (23.2)</b>	<b>+0.8 (9.3)</b>	<b>0.616</b>
		<b>[83.5 - 157.0]</b>	<b>[71.0 - 161.0]</b>	<b>[-2.4; +4.0]</b>	
HR <sub>ave</sub> (bpm)	Males (n=18)	109.1 (20.1)	110.3 (21.2)	1.2 (7.9)	0.522
	Females (n=17)	112.6 (16.8)	111.4 (17.2)	-1.3 (6.4)	0.420
	<b>Combined (n=35)</b>	<b>110.8 (18.4)</b>	<b>110.8 (19.1)</b>	<b>+0.0 (7.2)</b>	<b>0.998</b>
		<b>[83.5 - 157.0]</b>	<b>[75.0 - 151.0]</b>	<b>[-2.5; +2.5]</b>	
SaO <sub>2</sub> <sub>pre</sub> (%)	Males (n=18)	97.9 (1.0)	96.9 (1.6)	-1.0 (2.0)	0.046
	Females (n=17)	98.0 (1.1)	97.5 (1.6)	-0.5 (1.3)	0.187
	<b>Combined (n=35)</b>	<b>97.9 (1.0)</b>	<b>97.2 (1.7)</b>	<b>-0.5 (1)</b>	<b>0.073</b>
		<b>[95 - 100]</b>	<b>[94 - 100]</b>	<b>[-1.0; +0.5]</b>	
SaO <sub>2</sub> <sub>post</sub> (%)	Males (n=18)	97.7 (1.5)	96.7 (1.8)	-0.9 (2.3)	0.094
	Females (n=17)	97.0 (2.6)	97.5 (2.2)	0.5 (1.3)	0.135
	<b>Combined (n=35)</b>	<b>97.4 (2.1)</b>	<b>97.2 (2.0)</b>	<b>-0.2</b>	<b>0.552</b>
		<b>[89 - 100]</b>	<b>[91 - 100]</b>	<b>[-0.9; +0.5]</b>	

HR<sub>pre</sub>: Heart rate prior to 6MWT; HR<sub>post</sub>: Heart rate post 6MWT; HR<sub>ave</sub>: Average heart rate; SaO<sub>2</sub><sub>pre</sub>: Oxygen saturation prior to 6MWT; SaO<sub>2</sub><sub>post</sub>: Oxygen saturation post 6MWT

509

510

511

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

		ICC <sub>3,2</sub> [95% CI]	95% LOA	SEM	SEM%	MDC <sub>95</sub>	MDC <sub>95%</sub>
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
	<b>Combined (n=35)</b>	<b>0.97 [0.94 - 0.99]</b>	<b>-31.4; +42.7</b>	<b>13.7</b>	<b>2.3</b>	<b>37.8</b>	<b>6.3</b>
ICC: Intraclass correlation; 95% CI: 95% confidence interval; 95% LOA: 95% limit of agreements; SEM: Standard error of measurement; MDC <sub>95</sub> : Minimum detectable change at the 95% confidence interval							

513

Accepted copy