

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## Reliability, measurement error and minimum detectable change in mobility measures among community-dwelling adults aged 50 years and over

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-030475
Article Type:	Research
Date Submitted by the Author:	15-Mar-2019
Complete List of Authors:	Donoghue, Orna; University of Dublin Trinity College, The Irish Longitudinal Study on Ageing (TILDA) Savva, George; Quadram Institute Bioscience, Börsch-Supan, Axel; Max Planck Institute for Social Law and Social Policy, Munich Center for the Economics of Aging Kenny, RoseAnne; Trinity College Dublin, The Irish Longitudinal Study on Ageing (TILDA); St James Hospital, Mercer's Institute for Successful Ageing
Keywords:	repeatability, physical performance tests, longitudinal change, Epidemiology < TROPICAL MEDICINE

SCHOLARONE™  
Manuscripts

1  
2  
3 1 **Reliability, measurement error and minimum detectable change in mobility measures among**  
4  
5 2 **community-dwelling adults aged 50 years and over**  
6

7 3  
8  
9  
10 4 Orna A Donoghue, PhD <sup>a</sup>, George M Savva, PhD <sup>b</sup>, Axel Börsch-Supan, PhD <sup>c</sup>, Rose Anne Kenny, MD  
11  
12 5 <sup>a,d</sup>  
13

14 6  
15  
16  
17 7 **Affiliations:**

18  
19 8 <sup>a</sup> The Irish Longitudinal Study on Ageing (TILDA), Trinity College Dublin, Lincoln Place, Dublin 2,  
20  
21  
22 9 Ireland. Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie).

23  
24 10 <sup>b</sup> Quadram Institute Bioscience, Norwich Research Park, Norwich, UK. Email:  
25  
26  
27 11 [George.savva@quadram.ac.uk](mailto:George.savva@quadram.ac.uk).

28  
29 12 <sup>c</sup> Munich Center for the Economics of Aging, Max-Planck Institute for Social Law and Social Policy,  
30  
31  
32 13 Amalienst, Munich, Germany. Email: [axel@boersch-supan.de](mailto:axel@boersch-supan.de).

33  
34 14 <sup>a,d</sup> Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland. Email:  
35  
36 15 [rkenny@tcd.ie](mailto:rkenny@tcd.ie).

37  
38 16  
39  
40 17 **Corresponding author:** Dr Orna Donoghue, The Irish Longitudinal Study on Ageing (TILDA), Trinity  
41  
42  
43 18 College Dublin, Lincoln Place, Dublin 2, Ireland

44  
45 19 Tel: +353 1 896 4391; Fax: +353 1 896 2451; Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie)  
46  
47  
48 20

49  
50 21 **Acknowledgements:** The authors would like to acknowledge the contribution of the participants  
51  
52  
53 22 and members of the TILDA and SHARE teams.  
54

55 23  
56  
57 24 **Word count:** 3161  
58  
59  
60 25

## 1 Abstract

2 Objective: To examine the effects of repeat assessments, rater and time of day on reliability of  
3 mobility measures using a population-based sample of Irish adults aged  $\geq 50$  years.

4 Design: Test-retest reliability study.

5 Setting: Academic health assessment centre of The Irish Longitudinal Study on Ageing (TILDA).

6 Participants: 128 community-dwelling adults from the Survey for Health, Ageing and Retirement in  
7 Europe (SHARE) Ireland study who agreed to take part in the SHARE-Ireland / TILDA collaboration.

8 Interventions: Not applicable.

9 Outcome Measures: Participants performed Timed Up-and-Go (TUG), repeated chair stands (RCS)  
10 and walking speed tests administered by one of two raters. Repeat assessments were conducted  
11 1-4 months later. Participants were randomised with respect to a change in time (morning,  
12 afternoon) and whether or not the rater was changed between assessments. Within- and  
13 between-participant variance for each measure was estimated using mixed effects models. Intra-  
14 class correlation (ICC), standard error of measurement (SEM) and minimum detectable change  
15 (MDC) were reported.

16 Results: Average performance did not vary between baseline and repeat assessments in any test,  
17 except RCS. There were inter-rater effects for most tests ( $P < .001$ ) but limited time of day effects.

18 Reliability varied from ICC=0.66 (RCS) to ICC=0.88 (usual gait speed). MDC was 2.08 s for TUG, 4.52  
19 s for RCS and ranged from 19.49-34.73 cm/s for walking speed tests.

20 Conclusions: Reliability varied for each test when measurements are obtained over 1-4 months  
21 with most variation due to rater effects. Usual and motor dual task gait speed demonstrated  
22 highest reliability. MDC estimates provide guidance on whether longitudinal change in a similar  
23 group represents a genuine change in performance.

24 **Key words:** repeatability, physical performance tests, longitudinal change, epidemiology

## 1 Article summary

### 2 Strengths and limitations of this study

- 3 • This study provides information on the effects of repeat assessments, rater and time of day  
4 on reliability of mobility measures obtained over 1-4 months using a population-based  
5 sample of relatively healthy middle-aged and older aged  $\geq 50$  years in Ireland.
- 6 • The use of common tests such as Timed Up-and-Go, repeated chair stands and GAITRite  
7 assessments makes this analysis relevant for other studies looking at change in mobility.
- 8 • Mixed effects models were used to estimate within- and between-participant variance for  
9 each measure allowing intra-class correlation (ICC), standard error of measurement (SEM)  
10 and minimum detectable change (MDC) to be presented.
- 11 • For some measures, MDC was presented on the multiplicative (logarithmic) scale and the  
12 additive scale to account for skewness and to ensure that findings are applicable across all  
13 levels of performance. .

14  
15 **Funding:** TILDA received financial support from the Irish Government (Department of Health and  
16 Children), the Atlantic Philanthropies and Irish Life plc. The SHARE-TILDA project was funded by  
17 the National Institute of Aging (Prime Award Number R21AG040387). Funders had no involvement  
18 in analysis and preparation of this paper.

19  
20 **Competing interests:** None declared

21  
22 **Data sharing statement:** The anonymised SHARE-TILDA dataset is available through the on-site  
23 “hot desk” facility at TILDA, Trinity College Dublin. Researchers should contact [tilda@tcd.ie](mailto:tilda@tcd.ie) for  
24 more information.

## 1 Introduction

Performance based measures such as Timed Up-and-Go (TUG), repeated chair stands (RCS) and walking speed tests are commonly used to assess mobility and lower limb function of older adults in clinical and research settings<sup>1</sup>. These measures are good predictors of falls, disability, cognitive decline and mortality<sup>2-4</sup>. To be useful, they also need to be reliable (consistent when measured on several occasions and when there is no change in a subject's underlying performance) and responsive (able to detect a change when there is one)<sup>5</sup>. Good reliability allows changes in measurements to be tracked over time<sup>6</sup>.

However, all tests are subject to measurement error due to within-subject, inter-trial, inter-rater and day-to-day variation, and other external factors. This has several implications. Clinically, if an individual improves or declines between two testing sessions, it is important to know how likely it is that the observed change is a genuine change in performance and not due to measurement error. In research settings, unreliable measures can lead to regression dilution bias or false positive associations when testing predictors of longitudinal change<sup>7</sup>. To account for this, several measures of relative reliability i.e. intra-class correlation (ICC), and absolute reliability i.e. standard error of measurement (SEM) and minimum detectable change (MDC), are often reported<sup>8</sup>.

SEM is the standard deviation of the measurement error of a measure within an individual, for a given 'true' value of the underlying construct. The SEM determines the MDC, which is the smallest difference between two single observations that can be confidently attributed to a genuine difference and not measurement error. ICC is a measure of the proportion of variance within a population that is attributable to variance across individuals as opposed to measurement error within individuals. As opposed to SEM and MDC, ICC depends on both the SEM and the variation

1  
2  
3 1 between members of a sample, and so is not usually comparable or applicable across samples with  
4  
5 2 different levels of heterogeneity.

6  
7 3  
8  
9  
10 4 The within-session and one week test-retest reliability of TUG in community-dwelling, older adults  
11  
12 5 is well known, and is known to be high ( $ICC \geq 0.96$ )<sup>9-11</sup> in various populations as is the inter-rater<sup>12</sup>  
13  
14 6<sup>13</sup> and intra-rater reliability<sup>12</sup>. MDC at the 95% confidence level ( $MDC_{95}$ ) has been reported to vary  
15  
16  
17 7 between 3.33-6.87 s in healthy and cognitively impaired older adults<sup>14-16</sup> and up to 11 sec in  
18  
19 8 Parkinson's disease patients<sup>17</sup>. The within-session test-retest reliability of RCS is also very high  
20  
21  
22 9 ( $ICC=0.93-0.95$ )<sup>9,18</sup>, however SEM and MDC for community-dwelling adults are not known.

23  
24 10  
25  
26  
27 11 Walking speed can be measured using stopwatches, timing gates or sensed mats. The test-retest  
28  
29 12 reliability of usual gait speed (UGS) measured using a GAITRite® walkway has been reported to be  
30  
31  
32 13 between  $ICC=0.84$  and  $0.97$  for assessments given up to two weeks apart<sup>19-25</sup>. Similar values have  
33  
34 14 been reported for one hour test-retest reliability of dual task gait speed ( $ICC=0.85-0.93$ )<sup>19,20</sup>.  
35  
36  
37 15 Fewer studies have reported SEM or MDC in healthy populations with MDC values of 12.4-13.6  
38  
39 16 cm/s reported for UGS<sup>20,22</sup> and 15.5 cm/s for dual task gait speed<sup>20</sup>. However, reliability of dual  
40  
41  
42 17 task gait speed may also be dependent on the actual dual task and therefore is not comparable  
43  
44 18 across studies unless the same test has been used.

45  
46 19  
47  
48  
49 20 Here we report the test-retest reliability measured by ICC, SEM and MDC in a pragmatic  
50  
51 21 epidemiologic setting. We explore how reliability changes when lag between assessments varies  
52  
53  
54 22 between 1 month and 4 months, when rater changes or is held constant, and whether or not time  
55  
56 23 of assessment varies, in a large sample of healthy adults aged 50 and older recruited at random  
57  
58  
59  
60

1  
2  
3 1 from the population. This data will inform both clinical interpretation and design and analysis of  
4  
5 2 research studies using these tests.  
6  
7  
8 3

#### 10 4 **Methods**

##### 12 5 Participants

14 6 Participants were a subsample from the Survey of Health, Ageing and Retirement in Europe  
15  
16  
17 7 (SHARE), a longitudinal, cross-national study on health, socio-economic status and social and  
18  
19  
20 8 family networks of more than 80,000 individuals aged 50 years and over across Europe <sup>26</sup>. The  
21  
22 9 SHARE-Ireland sample (n=1,119) was recruited in Ireland between 2006 and 2007 with a response  
23  
24  
25 10 rate of 55% <sup>27</sup>. A collaboration between SHARE-Ireland and The Irish Longitudinal Study on Ageing  
26  
27 11 (TILDA) was established to understand the measurement properties of a comprehensive health  
28  
29  
30 12 assessment among a representative sample of the European population. Reliability of cognitive  
31  
32 13 measures and blood pressure dynamics based on this sample have been published previously <sup>28 29</sup>.

34 14  
35  
36 15 The extant SHARE-Ireland cohort at 2010 (n=827) was contacted and invited to take part in a  
37  
38  
39 16 health assessment delivered within the TILDA health assessment centre based at Trinity College  
40  
41  
42 17 Dublin. Initial contact was made by post and followed up by telephone between September 2011  
43  
44 18 and March 2012, with 377 participants consenting to receive further information about the study.  
45  
46 19 Of these, 253 agreed to an initial health assessment (see Figure 1). Ethical approval for this sub-  
47  
48  
49 20 study was obtained from the Faculty of Health Sciences Research Ethics Committee at Trinity  
50  
51 21 College Dublin. All participants provided informed consent.

##### 54 22 55 56 23 Health assessments and interview

57  
58  
59  
60



1  
2  
3 1 The full health assessment included a 3 hour battery of tests assessing cognitive function, gait and  
4  
5 2 mobility, cardiovascular function and vision<sup>30</sup>. Health assessments were conducted by two highly  
6  
7 3 trained research nurses with approximately 3 years' experience delivering these specific tests in  
8  
9  
10 4 the current setting. They used detailed and standardised health assessment protocols which  
11  
12 5 included clear explanations and demonstrations to ensure consistent instructions were provided  
13  
14 6 to all participants.  
15  
16

17 7  
18  
19 8 A short interview was administered by the nurses before the health assessment to capture  
20  
21  
22 9 information on health, chronic disease, disability, employment, social and financial circumstances.  
23  
24 10 Co-morbidity was assessed by asking participants if a doctor had ever told them that they had any  
25  
26  
27 11 of the following conditions: heart attack, high blood pressure, high cholesterol, stroke, diabetes,  
28  
29 12 chronic lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson's disease, cataracts,  
30  
31  
32 13 age related macular degeneration, Alzheimer's disease and atrial fibrillation. The number of  
33  
34 14 conditions was summed and categorised according to 0, 1, 2 or  $\geq 3$  conditions. Participants self-  
35  
36 15 rated their health as excellent, very good, good, fair or poor.  
37  
38  
39 16

40  
41 17 On completing the health assessment, 180 participants were invited to take part in an identical  
42  
43  
44 18 repeat assessment, scheduled after 1-4 months. In total, 128 participants (58 men) agreed to the  
45  
46 19 repeat assessment giving a response rate of 71% (25 refused and 27 were unavailable to attend  
47  
48  
49 20 the repeat assessment within the required timeframe).  
50

51 21  
52  
53 22 Repeat assessments were arranged to distinguish within-person variation from variation caused by  
54  
55  
56 23 changing rater or time of day. The same research nurse conducted the baseline and repeat  
57  
58 24 assessments for half of the participants while another nurse conducted the repeat assessment for  
59  
60

1  
2  
3 1 the other half of the participants. Time of day when the assessment took place (morning or  
4  
5 2 afternoon) was also changed for half of the participants. Change of rater, change of time of day  
6  
7 3 and delay between assessments (dichotomised at the median) were randomised using a  
8  
9  
10 4 minimisation routine designed to achieve balance across these covariates, as well as the age group  
11  
12 5 and sex of participants. Other factors that could influence performance e.g. health assessment  
13  
14 6 protocols, assessment location, equipment, etc. were held constant across both assessments.  
15  
16  
17 7

## 19 8 Physical performance tests

20  
21  
22 9 Participants completed several mobility tests - TUG, RCS and gait assessments in single and dual  
23  
24 10 task conditions. TUG, which is a common functional mobility test <sup>12</sup> was completed once using  
25  
26  
27 11 walking aids if required. The time taken to rise from the chair (seat height 46 cm), walk 3 m at  
28  
29 12 normal pace, turn around, walk back to the chair and sit down again was recorded using a  
30  
31  
32 13 stopwatch. RCS is an indicator of mobility and lower limb muscular endurance <sup>31</sup>. Participants  
33  
34 14 began in a seated position and the time taken to stand up five times was recorded. Participants  
35  
36  
37 15 were asked to keep their arms folded across their chest throughout the test.  
38  
39 16

40  
41 17 Gait assessment took place using a 4.88 m computerised walkway with embedded pressure  
42  
43  
44 18 sensors (GAITRite®, CIR Systems Inc, New York, USA). Participants performed two walks at their  
45  
46 19 normal pace followed by two walks under cognitive dual task conditions and manual dual task  
47  
48  
49 20 conditions. The cognitive task was to recite alternate letters of the alphabet (A-C-E, etc). The  
50  
51 21 manual task was to carry a glass of water filled to 7 mm from the top. Participants started and  
52  
53  
54 22 finished 2.5 m before and after the walkway to allow for acceleration and deceleration. The two  
55  
56 23 walks in each condition were combined to give mean UGS, mean cognitive dual task gait speed  
57  
58  
59 24 (CGS) and mean manual dual task gait speed (MGS).  
60

1  
2  
3 1 Statistical analysis

4  
5 2 This analysis includes participants who completed and had valid scores for baseline and repeat  
6  
7 3 assessments for each of the mobility tests (Figure 1). TUG and RCS are not normally distributed  
8  
9  
10 4 and the variance is strongly related to average scores, therefore analyses were conducted and  
11  
12 5 findings are presented on the natural scale for ease of interpretation and as log transformed  
13  
14 6 values to allow normally distributed stable variances across groups.  
15  
16

17 7  
18  
19 8 To look for practice effects, rater effects and time of day effects, mean mobility performance  
20  
21  
22 9 scores were compared (i) between baseline and repeat assessments, (ii) between raters, and (iii)  
23  
24 10 at different times of day using paired t-tests.  
25  
26

27 11  
28  
29 12 To estimate reliability, mixed effects regression models were then used to find the variation  
30  
31  
32 13 between and within participants. Baseline/repeat assessment, rater and time of day were included  
33  
34 14 as fixed effects. The standard deviations of the within-person and between-person variance  
35  
36  
37 15 components arising from these models were used to estimate the residual ICC for all measures  
38  
39 16 within this population. The ICC is the proportion of total variance not accounted for by within  
40  
41  
42 17 person variation, that is,  $ICC = \frac{SD_{Between}^2}{SD_{Between}^2 + SD_{Within}^2}$ . SEM is equivalent to  $SD_{Within}$ , the standard  
43  
44  
45 18 deviation of the variance of the test within individuals, assuming no genuine change in function,  
46  
47 19 and so is an absolute measure of test reliability. MDC is the magnitude of observable change  
48  
49  
50 20 required to exceed the anticipated measurement error and within-subject variability. It is  
51  
52 21 calculated by  $\sqrt{2} \times Z \times SD_{Within}$ , where  $Z=1.96$  for the 95% limit (that is, 95% of observed  
53  
54  
55 22 differences between pairs of observations will be within this limit given there is no true difference)  
56  
57 23 and  $Z=1.65$  for the 90% limit.  
58  
59  
60 24

1  
2 1 Findings from previous studies suggest that the variability of TUG is related to its magnitude; that  
3  
4  
5 2 is an individual with a TUG time of 4 s is likely to have a lower absolute variation than someone  
6  
7 3 with a TUG time of 12 s. For this reason, we estimate the reliability of TUG on a log-scale, as errors  
8  
9  
10 4 are more likely to be multiplicative than additive, and TUG is often analysed on a logarithmic scale  
11  
12 5 in epidemiological settings.  
13  
14  
15 6

## 17 7 Participant and public involvement

18  
19 8 This research was done without participant involvement. Participants were not invited to  
20  
21  
22 9 comment on the study design and were not consulted to develop participant relevant outcomes or  
23  
24 10 interpret the results. Participants were not invited to contribute to the writing or editing of this  
25  
26  
27 11 document for readability or accuracy.  
28  
29  
30 12

## 32 13 **Results**

33  
34 14 The median age of the sample was 66 years (range 51-89 years, IQR 61-71 years) and 55.5% were  
35  
36 15 female. The majority of the sample (n=103, 81.8%) rated their own health as excellent, very good  
37  
38  
39 16 or good, 57.8% reported having no history of cardiovascular or chronic conditions while 16.0% had  
40  
41 17 3 or more conditions. Median delay between assessments was 88 days (range 28-141 days, IQR  
42  
43  
44 18 70-104 days). Fifty-one participants had a different nurse at the repeat assessment while 60  
45  
46 19 participants had their assessment at a different time of day.  
47  
48  
49  
50 20

51  
52 21 Table 1 shows the mobility performance scores at baseline and repeat assessments, with different  
53  
54 22 raters and at different times of day, while Table 2 shows the variance and reliability estimates for  
55  
56  
57 23 all mobility measures.  
58  
59  
60 24

### 1 *Timed Up-and-go*

2 TUG did not vary between baseline and repeat assessments or by time of day, however there was  
3 a significant rater effect with a difference of 1.22 s ( $P<.001$ ) between the two nurses. The  
4 between-person SD was 1.31 s. The SEM was 0.75 s, leading to good reliability (ICC=0.75) and MDC  
5 estimates of 1.75 s at the 90% level and 2.08 s at the 95% level. This means that a difference of  
6 1.75-2.08 s between two assessments in the same individual can be expected by chance  
7 depending on the confidence interval used and when controlling for all other factors (rater, time  
8 between assessments and time of day). Analysis of TUG on a logarithmic scale suggests similar  
9 reliability (ICC=0.71), and a SEM of 0.09. The MDC<sub>95</sub> of 0.24 for log(TUG) suggests that change in  
10 TUG of up to 27% might be expected by chance in 95% of paired samples. This finding is applicable  
11 across the spectrum of baseline TUG scores.

### 12 *Repeated chair stands*

13 RCS was completed slightly more quickly at the repeat measurement (difference=0.47 s,  $P=.04$ )  
14 and when the assessment was carried out by Nurse 1 (difference=1.09 s,  $P<.001$ ) but did not vary  
15 with time of day. The ICC was 0.66 and SEM was 1.63 s while MDC was estimated to be 3.80 s at  
16 the 90% level and 4.52 s at the 95% level. Time to complete RCS was also analysed on the log  
17 scale, where reliability was similar (ICC=0.68), SEM was 0.13 and MDC was 0.35 at the 95%  
18 confidence level.

### 19 *Usual gait speed*

20 UGS did not vary between baseline and repeat assessment or by time of day, however there was a  
21 significant rater effect with a difference of 7.36 cm/s ( $P<.001$ ). Reliability was excellent (ICC=0.88)

1  
2  
3 1 as the between-person SD (18.65 cm/s) was much higher than the SEM (7.03 cm/s), resulting in a  
4  
5 2 MDC<sub>90</sub> of 16.40 cm/s and MDC<sub>95</sub> of 19.49 cm/s.  
6  
7 3

#### 10 4 *Manual dual task gait speed*

11  
12 5 Gait speed became less reliable as the complexity of the dual task conditions increased. MGS was  
13  
14 6 consistent across repeat assessments but varied by rater (difference=4.88 cm/s,  $P=.02$ ) and time of  
15  
16 7 day (difference=3.62 s,  $P=0.03$ ). ICC was lower than was observed for UGS (ICC=0.83), SEM was  
17  
18 8 higher (8.97 cm/s) and consequently so was MDC<sub>90</sub> (20.93 cm/s) and MDC<sub>95</sub> (24.87 cm/s).  
19  
20  
21  
22 9

#### 24 10 *Cognitive dual task gait speed*

25  
26  
27 11 CGS did not vary by repeat assessment, rater or time of day, however reliability estimates were  
28  
29 12 poorest out of all gait speed measures (ICC=0.77; SEM=12.53 cm/s; MDC<sub>95</sub>=34.73 cm/s).  
30  
31  
32 13

33  
34 14 For all observed rater effects, including those where performance was automatically measured  
35  
36 15 (i.e. with GAITRite), participants completed the mobility tasks more quickly when assessed by  
37  
38 16 Nurse 1.  
39  
40  
41 17

## 44 18 **Discussion**

45  
46 19 We report test-retest reliability, SEM and MDC of commonly used mobility tests in a sample of  
47  
48 20 relatively healthy, community-dwelling Irish adults aged 50 years and older. We found excellent  
49  
50 21 test-retest reliability for walking speed and motor dual task walking speed and good reliability for  
51  
52 22 TUG and cognitive dual task walking speed however, the lowest ICC was observed for RCS. These  
53  
54 23 findings contrast to previous studies which reported moderate to excellent reliability for all of  
55  
56 24 these measures<sup>9-11 18 19 21-25 32</sup>. As ICC depends on the distribution of scores within the sample it is  
57  
58  
59  
60

1  
2 1 estimated in and reflects relative reliability, it is specific to that particular setting and population <sup>8</sup>.  
3  
4  
5 2 Lower reliability here is likely to reflect more homogeneous population representative samples  
6  
7 3 (hence lower between-person standard deviations) compared to clinical samples with varying  
8  
9 4 degrees of impairment.  
10  
11

12 5  
13  
14 6 SEM and MDC provide an indication of absolute reliability. MDC allows the assessor to interpret if  
15  
16  
17 7 an observed change score is above that expected due to measurement error and therefore if it  
18  
19 8 represents a genuine change in performance. In this study, MDC for TUG (2.08 s at the 95% level)  
20  
21 9 is lower than that presented in previous studies of healthy (MDC<sub>95</sub>=4.71 s) <sup>16</sup> and cognitively  
22  
23 10 impaired (MDC<sub>95</sub>=5.88-6.87 s) older adults <sup>14 15</sup> and Parkinson's disease patients (MDC<sub>95</sub>=11 s) <sup>17</sup>.  
24  
25  
26 11 However, reporting variability in TUG as a percentage change in performance rather than in  
27  
28 12 absolute terms may be more appropriate. In contrast, MDC<sub>95</sub> for UGS, MGS and CGS  
29  
30 13 (MDC<sub>95</sub>=19.49-34.76 cm/s) are generally higher than the values estimated in community-dwelling  
31  
32 14 healthy adults (MDC<sub>95</sub>=13.6 cm/s) <sup>22</sup>, community-dwelling and hospitalised fallers (MDC<sub>95</sub>=12.4-  
33  
34 15 15.5 cm/s) <sup>32</sup> and in those post-stroke (MDC<sub>95</sub>=20 cm/s) <sup>33</sup>. These differences may be due to the  
35  
36 16 position on the performance scale as participants in these studies demonstrated poorer mobility  
37  
38 17 than participants in the SHARE-TILDA study <sup>22 32 33</sup>.  
39  
40  
41  
42  
43  
44  
45

46 19 Many longitudinal or intervention based studies vary widely in sample characteristics, co-  
47  
48 20 morbidity and time intervals between assessments. This makes cross-study comparisons difficult  
49  
50 21 and therefore reliability measures are best estimated for each sample and for groups with specific  
51  
52 22 diagnoses. This study provides guidance on MDC across the range of function in a generally  
53  
54 23 healthy, population-based sample, when measurements are compared weeks or months apart.  
55  
56 24 These estimates should be used when assessing individual changes in mobility performance over  
57  
58  
59  
60

1  
2  
3 1 this time-scale, when calculating required sample sizes for studies using these outcomes or  
4  
5 2 applying methods to adjust for measurement error in epidemiological studies. Participants in this  
6  
7 3 study were relatively healthy and so are unlikely to demonstrate a genuine change in performance  
8  
9  
10 4 in the time period examined.  
11

12 5  
13  
14 6 These results show the significant effect of inter-rater variation even with two highly trained and  
15  
16  
17 7 experienced research nurses. This suggests that changing rater introduces additional variance in  
18  
19 8 the measures beyond within-participant variation. The effect was observed in the GAITRite®  
20  
21  
22 9 assessment as well as stopwatch based tests suggesting that rater differences in reaction time do  
23  
24 10 not explain this. Both nurses were highly experienced and followed standardised protocols,  
25  
26  
27 11 however one explanation could be that they have different styles of interaction with respondents,  
28  
29 12 which may have impacted on the respondent's understanding of the task, or their motivation and  
30  
31  
32 13 subsequent desire to perform well. This emphasises the importance of providing appropriate  
33  
34 14 training for all raters to ensure that measurements are as accurate and consistent as possible.  
35  
36  
37 15 Where possible, analyses should be adjusted to account for differences between the raters  
38  
39 16 conducting the assessments.  
40

## 41 17 42 43 44 18 Study Strengths and Limitations

45  
46 19 A strength of this study is the population-based sample of relatively healthy middle-aged and older  
47  
48  
49 20 adults used in the analysis. In addition, our estimates of reliability remove time of day and rater  
50  
51 21 effects. For measures that are skewed, a different MDC may be required depending on whether  
52  
53  
54 22 performance is at the higher or lower ends of the spectrum. To account for this, we represent  
55  
56 23 relevant findings on the multiplicative (logarithmic) scale and the additive scale. Although a  
57  
58  
59 24 stopwatch is the easiest and most cost effective way to measure gait speed, the GAITRite® mat is  
60



1  
2 1 frequently used in research. Therefore, this analysis provides useful guidance on data obtained  
3  
4  
5 2 using simple and more complex instruments.  
6  
7 3  
8  
9

#### 10 4 **Conclusion**

11  
12 5 Gait speed obtained during normal walking conditions and when completing a manual dual task  
13  
14 6 are repeatable when performed at time intervals of several weeks to months, with lower reliability  
15  
16  
17 7 observed for the cognitive dual walk, TUG and RCS. There is also a potentially large effect of rater,  
18  
19 8 even for measures that are automatically measured. The estimates of MDC are presented for a  
20  
21 9 population based sample of relatively healthy middle-aged and older Irish adults and can be used  
22  
23  
24 10 to assess changes in performance in individuals drawn from comparable populations. Similar  
25  
26  
27 11 robust reliability studies are recommended to inform the use and interpretation of repeated  
28  
29 12 assessments in other populations such as those with specific co-morbidities. Additional analysis  
30  
31 13 using anchor-based approaches could be used to examine if these changes are of clinical  
32  
33  
34 14 importance.  
35  
36  
37 15  
38

#### 39 16 **Author contributions:**

40  
41 17 Substantial contributions to the conception or design of the work; or the acquisition, analysis, or  
42  
43 18 interpretation of data for the work – OD, GS, AB-S, RAK; Drafting the work or revising it critically  
44  
45  
46 19 for important intellectual content – OD, GS, AB-S, RAK; Final approval of the version to be  
47  
48  
49 20 published – OD, GS, AB-S, RAK; Agreement to be accountable for all aspects of the work in  
50  
51 21 ensuring that questions related to the accuracy or integrity of any part of the work are  
52  
53  
54 22 appropriately investigated and resolved – OD, GS, AB-S, RAK.  
55  
56 23  
57  
58 24  
59

#### 60 25 **References**

- 1  
2  
3 1 1. Kenny RA, Coen RF, Frewen J, et al. Normative values of cognitive and physical function in older  
4  
5 2 adults: findings from the Irish Longitudinal Study on Ageing. *Journal of the American*  
6  
7 3 *Geriatric Society* 2013;61(2):S279-S90.
- 8  
9  
10 4 2. Abellan Van Kan G, Rolland Y, Andrieu S, et al. Gait speed at usual pace as a predictor of adverse  
11  
12 5 outcomes in community-dwelling older people an International Academy on Nutrition and  
13  
14 6 Aging (IANA) Task Force. *Journal of Nutrition Health and Aging* 2009;13(10):881-89. doi:  
15  
16 7 10.1007/s12603-009-0246-z
- 17  
18  
19 8 3. Cooper R, Kuh D, Hardy R, et al. Objectively measured physical capability levels and mortality:  
20  
21 9 systematic review and meta-analysis. *BMJ* 2010;341(341):c4467. doi: 10.1136/bmj.c4467
- 22  
23  
24 10 4. Cooper R, Kuh D, Cooper C, et al. Objective measures of physical capability and subsequent  
25  
26 11 health: a systematic review. *Age Ageing* 2011;40(1):14-23.
- 27  
28  
29 12 5. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between  
30  
31 13 reproducibility and responsiveness. *Qual Life Res* 2001;10(7):571-8.
- 32  
33  
34 14 6. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30(1):1-  
35  
36 15 15.
- 37  
38  
39 16 7. Glymour MM, Weuve J, Berkman LF, et al. When is baseline adjustment useful in analyses of  
40  
41 17 change? An example with education and cognitive change. *Am J Epidemiol*  
42  
43 18 2005;162(3):267-78.
- 44  
45  
46 19 8. Rydwik E, Bergland A, Forsen L, et al. Investigation into the reliability and validity of the  
47  
48 20 measurement of elderly people's clinical walking speed: a systematic review. *Physiother*  
49  
50 21 *Theory Pract* 2012;28(3):238-56.
- 51  
52  
53 22 9. Griswold D, Rockwell K, Killa C, et al. Establishing the reliability and concurrent validity of  
54  
55 23 physical performance tests using virtual reality equipment for community-dwelling healthy  
56  
57 24 elders. *Disability and Rehabilitation* 2014;25:1-5. doi: doi:10.3109/09638288.2014.952451  
58  
59  
60

- 1  
2  
3 1 10. Regterschot GRH, Zhang W, Baldus H, et al. Test–retest reliability of sensor-based sit-to-stand  
4  
5 2 measures in young and older adults. *Gait Posture* 2014;40(1):220-24. doi:  
6  
7 3 <http://dx.doi.org/10.1016/j.gaitpost.2014.03.193>  
8  
9  
10 4 11. Ng SS, Hui-Chan CW. The Timed Up & Go Test: Its Reliability and Association With Lower-Limb  
11  
12 5 Impairments and Locomotor Capacities in People With Chronic Stroke. *Arch Phys Med*  
13  
14 6 *Rehabil* 2005;86(8):1641-47. doi: 10.1016/j.apmr.2005.01.011  
15  
16  
17 7 12. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail  
18  
19 8 elderly persons. *J Am Geriatr Soc* 1991;39(2):142-48. [published Online First: PMID:  
20  
21 9 1991946 ]  
22  
23  
24 10 13. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-  
25  
26 11 Dwelling Older Adults Using the Timed Up & Go Test. *Phys Ther* 2000;80(9):896-903.  
27  
28  
29 12 14. Blankevoort CG, van Heuvelen MJ, Scherder EJ. Reliability of six physical performance tests in  
30  
31 13 older people with dementia. *Phys Ther* 2013;93(1):69-78.  
32  
33  
34 14 15. Ries JD, Echternach JL, Nof L, et al. Test-retest reliability and minimal detectable change scores  
35  
36 15 for the timed "up & go" test, the six-minute walk test, and gait speed in people with  
37  
38 16 Alzheimer disease. *Phys Ther* 2009;89(6):569-79.  
39  
40  
41 17 16. Mangione KK, Craik RL, McCormick AA, et al. Detectable Changes in Physical Performance  
42  
43 18 Measures in Elderly African Americans. *Phys Ther* 2010;90(6):921-27. doi:  
44  
45 19 10.2522/ptj.20090363  
46  
47  
48  
49 20 17. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and  
50  
51 21 ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease  
52  
53 22 rating scale in people with parkinsonism. *Phys Ther* 2008;88(6):733-46.  
54  
55  
56 23 18. Goldberg A, Chavis M, Watkins J, et al. The five-times-sit-to-stand test: validity, reliability and  
57  
58 24 detectable change in older females. *Aging Clin Exp Res* 2012;24(4):339-44.  
59  
60

- 1  
2  
3 1 19. Hollman JH, Childs KB, McNeil ML, et al. Number of strides required for reliable measurements  
4  
5 2 of pace, rhythm and variability parameters of gait during normal and dual task walking in  
6  
7 3 older individuals. *Gait Posture* 2010;32(1):23-28. doi: 10.1016/j.gaitpost.2010.02.017  
8  
9  
10 4 20. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
11  
12 5 in community-dwelling and hospitalized older fallers. *Gait Posture* (38(4):1010-4)  
13  
14  
15 6 21. Brach JS, Perera S, Studenski S, et al. The Reliability and Validity of Measures of Gait Variability  
16  
17 7 in Community-Dwelling Older Adults. *Arch Phys Med Rehabil* 2008;89(12):2293-96. doi:  
18  
19 8 10.1016/j.apmr.2008.06.010  
20  
21  
22 9 22. Goldberg A, Schepens S. Measurement error and minimum detectable change in 4-meter gait  
23  
24 10 speed in older adults. *Aging Clin Exp Res* 2011;23(5-6):406-12.  
25  
26  
27 11 23. Menz HB, Lord SR, St George R, et al. Walking stability and sensorimotor function in older  
28  
29 12 people with diabetic peripheral neuropathy. *Arch Phys Med Rehabil* 2004;85(2):245-52.  
30  
31 13 doi: 10.1016/j.apmr.2003.06.015  
32  
33  
34 14 24. Van Iersel MB, Benraad CEM, Olde Rikkert MGM. Validity and reliability of quantitative gait  
35  
36 15 analysis in geriatric patients with and without dementia. *J Am Geriatr Soc*  
37  
38 16 2007;55(4):632-34. doi: 10.1111/j.1532-5415.2007.01130.x  
39  
40  
41 17 25. Paterson KL, Hill KD, Lythgo ND, et al. The reliability of spatiotemporal gait data for young and  
42  
43 18 older women during continuous overground walking. *Arch Phys Med Rehabil*  
44  
45 19 2008;89(12):2360-5.  
46  
47  
48  
49 20 26. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: The Survey of Health,  
50  
51 21 Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*  
52  
53 22 2013;42(4):992-1001.  
54  
55  
56 23 27. UCD Geary Institute and Irish Centre for Social Gerontology N. SHARE Ireland - Survey of  
57  
58 24 Health, Ageing and Retirement in Europe [Internet]. *Available from:*  
59  
60

- 1  
2  
3 1 [http://gearyucdie/share/fileadmin/user\\_upload/shareresults/Share\\_Wave1\\_Resultspdf](http://gearyucdie/share/fileadmin/user_upload/shareresults/Share_Wave1_Resultspdf)  
4  
5 2 2008  
6  
7 3 28. Feeney J, Savva GM, O'Regan C, et al. Measurement Error, Reliability, and Minimum  
8  
9  
10 4 Detectable Change in the Mini-Mental State Examination, Montreal Cognitive Assessment,  
11  
12 5 and Color Trails Test among Community Living Middle-Aged and Older Adults. *J Alzheimers*  
13  
14 6 *Dis* 2016;53(3):1107-14. doi: 10.3233/jad-160248 [published Online First: 2016/06/04]  
15  
16  
17 7 29. Finucane C, Savva GM, Kenny RA. Reliability of orthostatic beat-to-beat blood pressure tests:  
18  
19 8 implications for population and clinical studies. *Clin Auton Res* 2017;27(1):31-39. doi:  
20  
21 9 10.1007/s10286-016-0393-3 [published Online First: 2017/01/14]  
22  
23  
24 10 30. Cronin H, O'Regan C, Finucane C, et al. Health and aging: development of the Irish Longitudinal  
25  
26 11 Study on Ageing health assessment. *J Am Geriatr Soc* 2013;61(2):12197.  
27  
28  
29 12 31. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing  
30  
31 13 lower extremity function: association with self-reported disability and prediction of  
32  
33 14 mortality and nursing home admission. *J Gerontol* 1994;49(2):M85-94.  
34  
35  
36 15 32. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
37  
38 16 in community-dwelling and hospitalized older fallers. *Gait Posture* 2013;38(4):1010-4.  
39  
40  
41 17 33. Lewek MD, Randall EP. Reliability of spatiotemporal asymmetry during overground walking for  
42  
43 18 individuals following chronic stroke. *J Neurol Phys Ther* 2011;35(3):116-21.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 1. Mobility performance scores obtained at baseline and repeat assessments, with different raters and at different times of day.

	Assessment		Rater <sup>a</sup>		Time of day <sup>b</sup>	
	Baseline	Repeat	Nurse 1	Nurse 2	Test AM	Test PM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
TUG (s)	8.88 (1.39)	8.87 (1.54)	8.13 (1.20)	9.35 (1.51)***	8.83 (1.49)	8.69 (1.25)
log(TUG)	2.17 (0.02)	2.17 (0.01)	2.08 (0.02)	2.22 (0.02)***	2.16 (0.02)	2.15 (0.02)
RCS (s)	12.49 (2.87)	12.02 (2.48)*	11.80 (2.27)	12.89 (2.88)***	12.17 (2.99)	12.00 (2.46)
logRCS	2.50 (0.22)	2.46 (0.21)*	2.45 (0.20)	2.53 (0.24)**	2.47 (0.24)	2.46 (0.22)
UGS (cm/s)	137.95 (20.21)	138.20 (19.32)	145.82 (18.94)	138.46 (17.85)***	137.62 (17.68)	137.74 (17.38)
MGS (cm/s)	116.76 (21.84)	118.71 (19.93)	123.07 (18.95)	118.07 (20.45)**	117.86 (19.85)	122.19 (17.21)*
CGS (cm/s)	115.23 (24.08)	115.15 (25.21)	118.29 (25.24)	117.40 (20.99)	117.45 (24.01)	118.84 (20.18)

Notes: TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task gait speed

<sup>a</sup> Rater scores are calculated only among participants who changed rater at the repeat assessment

<sup>b</sup> Time of day scores are calculated only among participants who changed time of day at the repeat assessment.

\*  $P < 0.05$ ; \*\*  $P < 0.01$ ; \*\*\*  $P < 0.001$

Table 2. Variance and reliability estimates for all mobility tests.

	<b>SD<sub>between</sub> (95% CI)</b>	<b>SEM (95% CI)</b>	<b>ICC (95% CI)</b>	<b>MDC<sub>90</sub></b>	<b>MDC<sub>95</sub></b>
TUG (s)	1.31 (1.12-1.52)	0.75 (0.66-0.85)	0.75 (0.66-0.82)	1.75	2.08
logTUG	0.13 (0.11-0.15)	0.09 (0.08-0.10)	0.71 (0.61-0.79)	0.2	0.24
RCS (s)	2.29 (1.93-2.70)	1.63 (1.43-1.86)	0.66 (0.55-0.76)	3.8	4.52
logRCS	0.18 (0.16-0.22)	0.13 (0.11-0.14)	0.68 (0.57-0.77)	0.29	0.35
UGS (cm/s)	18.65 (16.34-21.29)	7.03 (6.20-7.98)	0.88 (0.83-0.91)	16.4	19.49
MGS (cm/s)	19.57 (17.04-22.46)	8.97 (7.90-10.19)	0.83 (0.76-0.88)	20.93	24.87
CGS (cm/s)	22.73 (19.62-26.34)	12.53 (10.99-14.28)	0.77 (0.68-0.83)	29.24	34.73

Notes: SEM, standard error of the measurement; TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task speed; ICC, intra-class correlation; MDC, minimum detectable change

Figure 1: Flow chart of participation in the study

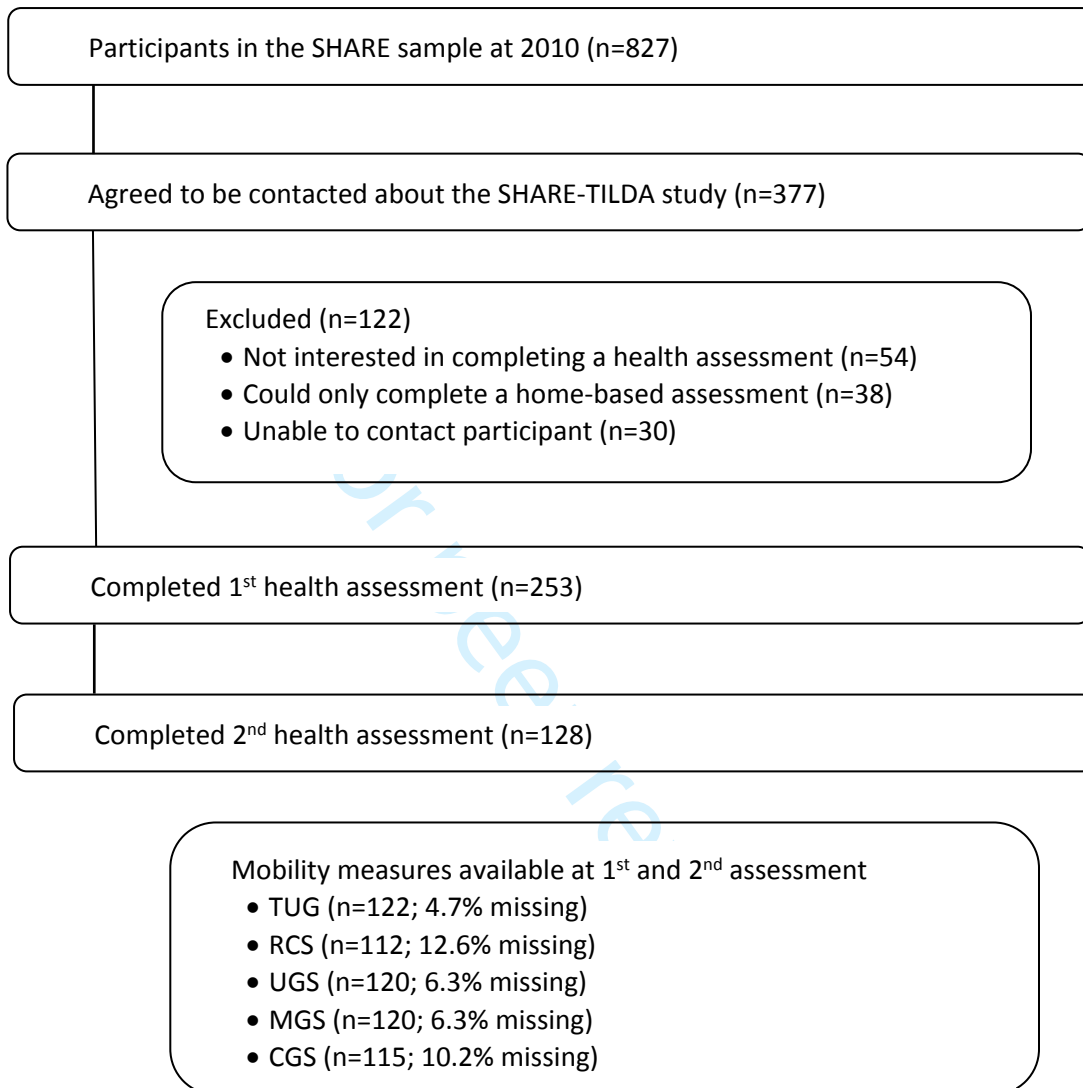


Figure 1: Exclusion criteria used to establish eligible participants for this analysis.

Note: CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair stands; SHARE, Survey for Health Ageing and Retirement in Europe; TILDA, The Irish Longitudinal Study on Ageing; TUG, Timed Up-and-Go; UGS, usual gait speed.



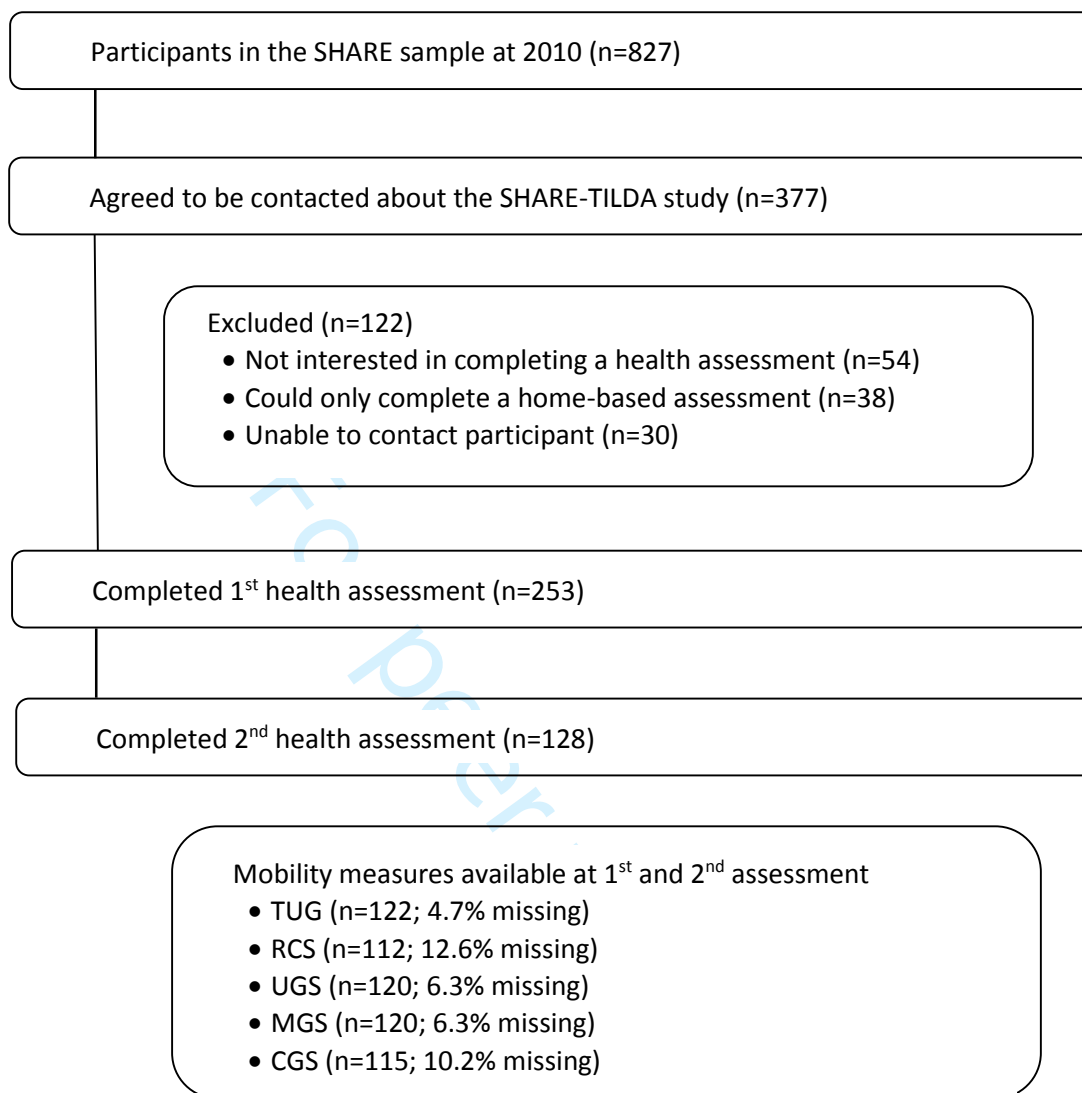


Figure 1: Exclusion criteria used to establish eligible participants for this analysis.

Note: CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair stands; SHARE, Survey for Health Ageing and Retirement in Europe; TILDA, The Irish Longitudinal Study on Ageing; TUG, Timed Up-and-Go; UGS, usual gait speed.

# BMJ Open

## Reliability, measurement error and minimum detectable change in mobility measures: a cohort study of community-dwelling adults aged 50 years and over in Ireland

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-030475.R1
Article Type:	Original research
Date Submitted by the Author:	23-Jul-2019
Complete List of Authors:	Donoghue, Orna; University of Dublin Trinity College, The Irish Longitudinal Study on Ageing (TILDA) Savva, George; Quadram Institute Bioscience, Börsch-Supan, Axel; Max Planck Institute for Social Law and Social Policy, Munich Center for the Economics of Aging Kenny, RoseAnne; Trinity College Dublin, The Irish Longitudinal Study on Ageing (TILDA); St James Hospital, Mercer's Institute for Successful Ageing
<b>Primary Subject Heading</b>:	Geriatric medicine
Secondary Subject Heading:	Geriatric medicine
Keywords:	repeatability, physical performance tests, longitudinal change, Epidemiology < TROPICAL MEDICINE

SCHOLARONE™  
Manuscripts

1  
2  
3 1 **Reliability, measurement error and minimum detectable change in mobility measures: a cohort**  
4  
5 2 **study of community-dwelling adults aged 50 years and over in Ireland**  
6

7 3  
8  
9  
10 4 Orna A Donoghue, PhD <sup>a</sup>, George M Savva, PhD <sup>b</sup>, Axel Börsch-Supan, PhD <sup>c</sup>, Rose Anne Kenny, MD  
11  
12 5 <sup>a,d</sup>  
13

14 6  
15  
16  
17 7 **Affiliations:**

18  
19 8 <sup>a</sup> The Irish Longitudinal Study on Ageing (TILDA), Trinity College Dublin, Lincoln Place, Dublin 2,  
20  
21  
22 9 Ireland. Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie).  
23

24 10 <sup>b</sup> Quadram Institute Bioscience, Norwich Research Park, Norwich, UK. Email:  
25  
26  
27 11 [George.savva@quadram.ac.uk](mailto:George.savva@quadram.ac.uk).  
28

29 12 <sup>c</sup> Munich Center for the Economics of Aging, Max-Planck Institute for Social Law and Social Policy,  
30  
31  
32 13 Amalienst, Munich, Germany. Email: [axel@boersch-supan.de](mailto:axel@boersch-supan.de).  
33

34 14 <sup>a,d</sup> Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland. Email:  
35  
36 15 [rkenny@tcd.ie](mailto:rkenny@tcd.ie).  
37

38 16  
39  
40 17 **Corresponding author:** Dr Orna Donoghue, The Irish Longitudinal Study on Ageing (TILDA), Trinity  
41  
42  
43 18 College Dublin, Lincoln Place, Dublin 2, Ireland  
44  
45 19 Tel: +353 1 896 4391; Fax: +353 1 896 2451; Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie)  
46  
47  
48 20

49  
50 21 **Acknowledgements:** The authors would like to acknowledge the contribution of the participants  
51  
52  
53 22 and members of the TILDA and SHARE teams.  
54

55 23  
56  
57 24 **Word count:** 3161  
58  
59  
60 25

## 1 Abstract

2 Objective: To examine the effects of repeat assessments, rater and time of day on reliability of  
3 mobility measures using a population-based sample of Irish adults aged  $\geq 50$  years.

4 Design: Test-retest cohort reliability study.

5 Setting: Academic health assessment centre of The Irish Longitudinal Study on Ageing (TILDA).

6 Participants: 128 community-dwelling adults from the Survey for Health, Ageing and Retirement in  
7 Europe (SHARE) Ireland study who agreed to take part in the SHARE-Ireland / TILDA collaboration.

8 Interventions: Not applicable.

9 Outcome Measures: Participants performed Timed Up-and-Go (TUG), repeated chair stands (RCS)  
10 and walking speed tests administered by one of two raters. Repeat assessments were conducted  
11 1-4 months later. Participants were randomised with respect to a change in time (morning,  
12 afternoon) and whether or not the rater was changed between assessments. Within- and  
13 between-participant variance for each measure was estimated using mixed effects models. Intra-  
14 class correlation (ICC), standard error of measurement (SEM) and minimum detectable change  
15 (MDC) were reported.

16 Results: Average performance did not vary between baseline and repeat assessments in any test,  
17 except RCS. There were inter-rater effects for most tests ( $P < .001$ ) but limited time of day effects.

18 Reliability varied from ICC=0.66 (RCS) to ICC=0.88 (usual gait speed). MDC was 2.08 s for TUG, 4.52  
19 s for RCS and ranged from 19.49-34.73 cm/s for walking speed tests.

20 Conclusions: Reliability varied for each test when measurements are obtained over 1-4 months  
21 with most variation due to rater effects. Usual and motor dual task gait speed demonstrated  
22 highest reliability. MDC estimates provide guidance on whether longitudinal change in a similar  
23 group represents a genuine change in performance.

24 **Key words:** repeatability, physical performance tests, longitudinal change, epidemiology

## 1 Article summary

### 2 Strengths and limitations of this study

- 3 • This study provides information on the effects of repeat assessments, rater and time of day  
4 on test-retest reliability of mobility measures obtained over 1-4 months using a population-  
5 based sample of relatively healthy middle-aged and older aged  $\geq 50$  years in Ireland.
- 6 • The use of common tests such as Timed Up-and-Go, repeated chair stands and GAITRite  
7 assessments makes this analysis relevant for other studies looking at change in mobility.
- 8 • Mixed effects models were used to estimate within- and between-participant variance for  
9 each measure allowing intra-class correlation (ICC), standard error of measurement (SEM)  
10 and minimum detectable change (MDC) to be presented.
- 11 • For some measures, MDC was presented on the multiplicative (logarithmic) scale and the  
12 additive scale to account for skewness and to ensure that findings are applicable across all  
13 levels of performance.
- 14 • Changes in exercise levels, activities, medications and current injury status could contribute  
15 to measurement variation.

1  
2  
3 1 **Funding:** TILDA received financial support from the Irish Government (Department of Health and  
4  
5 2 Children), the Atlantic Philanthropies and Irish Life plc. The SHARE-TILDA project was funded by  
6  
7 3 the National Institute of Aging (Prime Award Number R21AG040387). Funders had no involvement  
8  
9  
10 4 in analysis and preparation of this paper.  
11  
12 5

13  
14 6 **Competing interests:** None declared  
15  
16  
17 7

18  
19 8 **Data sharing statement:** The anonymised SHARE-TILDA dataset is available through the on-site  
20  
21  
22 9 “hot desk” facility at TILDA, Trinity College Dublin. Researchers should contact [tilda@tcd.ie](mailto:tilda@tcd.ie) for  
23  
24 10 more information.  
25  
26  
27 11  
28  
29 12  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 1 Introduction

Performance based measures such as Timed Up-and-Go (TUG), repeated chair stands (RCS) and walking speed tests are commonly used to assess mobility and lower limb function of older adults in clinical and research settings<sup>1</sup>. These measures are good predictors of falls, disability, cognitive decline and mortality<sup>2-4</sup>. To be useful, they also need to be reliable (consistent when measured on several occasions and when there is no change in a subject's underlying performance) and responsive (able to detect a change when there is one)<sup>5</sup>. Good reliability allows changes in measurements to be tracked over time<sup>6</sup>.

However, all tests are subject to measurement error due to within-subject, inter-trial, inter-rater and day-to-day variation, and other external factors. This has several implications. Clinically, if an individual improves or declines between two testing sessions, it is important to know how likely it is that the observed change is a genuine change in performance and not due to measurement error. In research settings, unreliable measures can lead to regression dilution bias or false positive associations when testing predictors of longitudinal change<sup>7</sup>. To account for this, several measures of relative reliability i.e. intra-class correlation (ICC), and absolute reliability i.e. standard error of measurement (SEM) and minimum detectable change (MDC), are often reported<sup>8</sup>.

SEM is the standard deviation of the measurement error of a measure within an individual, for a given 'true' value of the underlying construct. The SEM determines the MDC, which is the smallest difference between two single observations that can be confidently attributed to a genuine difference and not measurement error. ICC is a measure of the proportion of variance within a population that is attributable to variance across individuals as opposed to measurement error within individuals. As opposed to SEM and MDC, ICC depends on both the SEM and the variation

1  
2  
3 1 between members of a sample, and so is not usually comparable or applicable across samples with  
4  
5 2 different levels of heterogeneity.  
6

7  
8 3  
9  
10 4 The within-session and one week test-retest reliability of TUG in community-dwelling, older adults  
11  
12 5 is well known, and is known to be high ( $ICC \geq 0.96$ )<sup>9-11</sup> in various populations as is the inter-rater<sup>12</sup>  
13  
14 6<sup>13</sup> and intra-rater reliability<sup>12</sup>. MDC at the 95% confidence level ( $MDC_{95}$ ) has been reported to vary  
15  
16  
17 7 between 3.33-6.87 s in healthy and cognitively impaired older adults<sup>14-16</sup> and up to 11 sec in  
18  
19 8 Parkinson's disease patients<sup>17</sup>. The within-session test-retest reliability of RCS is also very high  
20  
21  
22 9 ( $ICC=0.93-0.95$ )<sup>9,18</sup>, however SEM and MDC for community-dwelling adults are not known.  
23

24  
25 10  
26  
27 11 Walking speed can be measured using stopwatches, timing gates or sensed mats. The test-retest  
28  
29 12 reliability of usual gait speed (UGS) measured using a GAITRite® walkway has been reported to be  
30  
31  
32 13 between  $ICC=0.84$  and  $0.97$  for assessments given up to two weeks apart<sup>19-25</sup>. Similar values have  
33  
34 14 been reported for one hour test-retest reliability of dual task gait speed ( $ICC=0.85-0.93$ )<sup>19,20</sup>.  
35  
36  
37 15 Fewer studies have reported SEM or MDC in healthy populations with MDC values of 12.4-13.6  
38  
39 16 cm/s reported for UGS<sup>20,22</sup> and 15.5 cm/s for dual task gait speed<sup>20</sup>. However, reliability of dual  
40  
41  
42 17 task gait speed may also be dependent on the actual dual task and therefore is not comparable  
43  
44 18 across studies unless the same test has been used.  
45

46  
47 19  
48  
49 20 Here we report the test-retest reliability measured by ICC, SEM and MDC in a pragmatic  
50  
51 21 epidemiologic setting. We explore how reliability changes when lag between assessments varies  
52  
53  
54 22 between 1 month and 4 months, when rater changes or is held constant, and whether or not time  
55  
56 23 of assessment varies, in a large sample of healthy adults aged 50 and older recruited at random  
57  
58  
59 24 from the population.  
60



1  
2  
3 1 In epidemiologic settings, the measures we have tested are commonly used as proxies for the  
4  
5 2 underlying general cognitive and physical health status of participants around the time of the  
6  
7 3 assessment. Short-term fluctuations in these measures, for example due to acute illness or day-to-  
8  
9 4 day variation, add error to these outcomes along with measurement error associated with the  
10  
11 5 instruments themselves. Hence when comparing measures over longer time periods, that is, years  
12  
13 6 or decades typical of epidemiologic research, it is important to know how well single measures of  
14  
15 7 physical and cognitive function reflect the underlying health status of the participant, net of any  
16  
17 8 factors that might cause a short-term fluctuation. Therefore, we tested the concordance between  
18  
19 9 pairs of measures between one and four months apart, to estimate the error association with both  
20  
21 10 measurement and short-term variation in each measure. Understanding natural variation in  
22  
23 11 outcomes over one to four months is also essential when planning clinical trials with follow-up  
24  
25 12 time in this range.  
26  
27  
28  
29  
30  
31  
32  
33

## 34 14 **Methods**

### 35 15 ***Participants***

36  
37 16 Participants were a subsample from the Survey of Health, Ageing and Retirement in Europe  
38  
39 17 (SHARE), a longitudinal, cross-national study on health, socio-economic status and social and  
40  
41 18 family networks of more than 80,000 individuals aged 50 years and over across Europe <sup>26</sup>. The  
42  
43 19 SHARE-Ireland sample (n=1,119) was recruited in Ireland between 2006 and 2007 with a response  
44  
45 20 rate of 55% <sup>27</sup>. A collaboration between SHARE-Ireland and The Irish Longitudinal Study on Ageing  
46  
47 21 (TILDA) was established to understand the measurement properties of a comprehensive health  
48  
49 22 assessment among a representative sample of the European population. Reliability of cognitive  
50  
51 23 measures and blood pressure dynamics based on this sample have been published previously <sup>28 29</sup>.  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 1 The extant SHARE-Ireland cohort at 2010 (n=827) was contacted and invited to take part in a  
4  
5 2 health assessment that included the same tests and followed the same protocols as those used by  
6  
7 3 TILDA. The health assessment was delivered to the SHARE-Ireland participants by TILDA research  
8  
9  
10 4 nurses within the TILDA health assessment centre based at Trinity College Dublin. Initial contact  
11  
12 5 was made by post and followed up by telephone between September 2011 and March 2012, with  
13  
14 6 377 participants consenting to receive further information about the study. Of these, 253 agreed  
15  
16  
17 7 to an initial health assessment (see Figure 1). Ethical approval for this sub-study was obtained  
18  
19  
20 8 from the Faculty of Health Sciences Research Ethics Committee at Trinity College Dublin. All  
21  
22 9 participants provided informed consent.  
23  
24  
25  
26

### 27 11 ***Health assessments and interview***

28  
29 12 The full health assessment included a 3 hour battery of tests assessing cognitive function, gait and  
30  
31 13 mobility, cardiovascular function and vision<sup>30</sup>. Health assessments were conducted by two highly  
32  
33  
34 14 trained research nurses with approximately 3 years' experience delivering these specific tests in  
35  
36  
37 15 the current setting. Training took approximately one month and nurses used detailed and  
38  
39 16 standardised health assessment protocols which included clear explanations and demonstrations  
40  
41  
42 17 to ensure consistent instructions were provided to all participants. Nurses also underwent periodic  
43  
44 18 quality control procedures to ensure adherence to the protocols.  
45  
46  
47  
48

49 20 A short interview was administered by the nurses before the health assessment to capture  
50  
51 21 information on health, chronic disease, disability, employment, social and financial circumstances.  
52  
53  
54 22 Co-morbidity was assessed by asking participants if a doctor had ever told them that they had any  
55  
56 23 of the following conditions: heart attack, high blood pressure, high cholesterol, stroke, diabetes,  
57  
58  
59 24 chronic lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson's disease, cataracts,  
60

1  
2 1 age related macular degeneration, Alzheimer's disease and atrial fibrillation. The number of  
3  
4  
5 2 conditions was summed and categorised according to 0, 1, 2 or  $\geq 3$  conditions. Participants self-  
6  
7 3 rated their health as excellent, very good, good, fair or poor.  
8  
9

10 4  
11  
12 5 On completing the health assessment, 180 participants were invited to take part in an identical  
13  
14 6 repeat assessment, scheduled after 1-4 months. In total, 128 participants (58 men) agreed to the  
15  
16  
17 7 repeat assessment giving a response rate of 71% (25 refused and 27 were unavailable to attend  
18  
19 8 the repeat assessment within the required timeframe).  
20  
21

22 9  
23  
24 10 Repeat assessments were arranged to distinguish within-person variation from variation caused by  
25  
26  
27 11 changing rater or time of day. The same research nurse conducted the baseline and repeat  
28  
29 12 assessments for half of the participants while another nurse conducted the repeat assessment for  
30  
31 13 the other half of the participants. Time of day when the assessment took place (morning or  
32  
33  
34 14 afternoon) was also changed for half of the participants. Change of rater, change of time of day  
35  
36 15 and delay between assessments (dichotomised at the median) were randomised using a  
37  
38  
39 16 minimisation routine designed to achieve balance across these covariates, as well as the age group  
40  
41 17 and sex of participants. Other factors that could influence performance e.g. health assessment  
42  
43  
44 18 protocols, assessment location, equipment, etc. were held constant across both assessments.  
45  
46

### 47 19 48 49 20 ***Physical performance tests***

50  
51 21 Participants completed several mobility tests - TUG, RCS and gait assessments in single and dual  
52  
53 22 task conditions. TUG, which is a common functional mobility test <sup>12</sup> was completed once using  
54  
55  
56 23 walking aids if required. The time taken to rise from the chair (seat height 46 cm), walk 3 m at  
57  
58  
59 24 normal pace, turn around, walk back to the chair and sit down again was recorded using a  
60

1  
2 1 stopwatch. RCS is an indicator of mobility and lower limb muscular endurance<sup>31</sup>. Participants  
3  
4  
5 2 began in a seated position and the time taken to stand up five times was recorded. Participants  
6  
7 3 were asked to keep their arms folded across their chest throughout the test.  
8  
9  
10 4

11  
12 5 Gait assessment took place using a 4.88 m computerised walkway with embedded pressure  
13  
14 6 sensors (GAITRite®, CIR Systems Inc, New York, USA). Participants performed two walks at their  
15  
16 7 normal pace followed by two walks under cognitive dual task conditions and manual dual task  
17  
18 8 conditions. The cognitive task was to recite alternate letters of the alphabet (A-C-E, etc). The  
19  
20 9 manual task was to carry a glass of water filled to 7 mm from the top. Participants started and  
21  
22 10 finished 2.5 m before and after the walkway to allow for acceleration and deceleration. The two  
23  
24 11 walks in each condition were combined to give mean UGS, mean cognitive dual task gait speed  
25  
26 12 (CGS) and mean manual dual task gait speed (MGS).  
27  
28  
29  
30  
31  
32  
33

#### 34 14 **Statistical analysis**

35  
36 15 This analysis includes participants who completed and had valid scores for baseline and repeat  
37  
38 16 assessments for each of the mobility tests (Figure 1). Missing data was not imputed. TUG and RCS  
39  
40 17 are not normally distributed and the variance is strongly related to average scores, therefore  
41  
42 18 analyses were conducted and findings are presented on the natural scale for ease of interpretation  
43  
44 19 and as log transformed values to allow normally distributed stable variances across groups.  
45  
46  
47  
48  
49 20

50  
51 21 To look for practice effects, rater effects and time of day effects, mean mobility performance  
52  
53 22 scores were compared (i) between baseline and repeat assessments, (ii) between raters, and (iii)  
54  
55 23 at different times of day using paired t-tests.  
56  
57  
58  
59 24  
60

1  
2  
3 1 To estimate reliability, mixed effects regression models were then used to find the variation  
4  
5 2 between and within participants. Baseline/repeat assessment, rater and time of day were included  
6  
7 3 as fixed effects. The standard deviations of the within-person and between-person variance  
8  
9  
10 4 components arising from these models were used to estimate the residual ICC for all measures  
11  
12 5 within this population. The ICC is the proportion of total variance not accounted for by within  
13  
14  
15 6 person variation, that is,  $ICC = \frac{SD_{Between}^2}{SD_{Between}^2 + SD_{Within}^2}$ . Koo & Li<sup>32</sup> recommend that the 95% confidence  
16  
17  
18 7 interval of the ICC estimate is used to evaluate reliability and suggest the following guidelines: <0.5  
19  
20  
21 8 indicates poor reliability, 0.5-0.75 indicates moderate reliability, 0.75-0.90 indicates good  
22  
23 9 reliability, >0.90 indicates excellent reliability.

24  
25  
26  
27  
28 11 SEM is equivalent to  $SD_{Within}$  the standard deviation of the variance of the test within individuals,  
29  
30 12 assuming no genuine change in function, and so is an absolute measure of test reliability. MDC is  
31  
32  
33 13 the magnitude of observable change required to exceed the anticipated measurement error and  
34  
35 14 within-subject variability. It is calculated by  $\sqrt{2} \times Z \times SD_{Within}$  where  $Z=1.96$  for the 95% limit  
36  
37  
38 15 (that is, 95% of observed differences between pairs of observations will be within this limit given  
39  
40 16 there is no true difference) and  $Z=1.65$  for the 90% limit.

41  
42  
43  
44  
45 18 Findings from previous studies suggest that the variability of TUG is related to its magnitude; that  
46  
47  
48 19 is an individual with a TUG time of 4 s is likely to have a lower absolute variation than someone  
49  
50 20 with a TUG time of 12 s. For this reason, we estimate the reliability of TUG on a log-scale, as errors  
51  
52  
53 21 are more likely to be multiplicative than additive, and TUG is often analysed on a logarithmic scale  
54  
55 22 in epidemiological settings.

## 23 24 **Participant and public involvement**

1  
2  
3 1 This research was done without participant involvement. Participants were not invited to  
4  
5 2 comment on the study design and were not consulted to develop participant relevant outcomes or  
6  
7 3 interpret the results. Participants were not invited to contribute to the writing or editing of this  
8  
9  
10 4 document for readability or accuracy.  
11  
12 5

## 14 6 **Results**

17 7 The median age of the sample was 66 years (range 51-89 years, IQR 61-71 years) and 55.5% were  
18  
19 8 female. The majority of the sample (n=103, 81.8%) rated their own health as excellent, very good  
20  
21  
22 9 or good, 57.8% reported having no history of cardiovascular or chronic conditions while 16.0% had  
23  
24 10 3 or more conditions. Median delay between assessments was 88 days (range 28-141 days, IQR  
25  
26  
27 11 70-104 days). Fifty-one participants had a different nurse at the repeat assessment while 60  
28  
29 12 participants had their assessment at a different time of day.  
30  
31  
32  
33 13

35 14 Table 1 shows the mobility performance scores at baseline and repeat assessments, with different  
36  
37 15 raters and at different times of day, while Table 2 shows the variance and reliability estimates for  
38  
39  
40 16 all mobility measures. In general, this sample was relatively robust with good levels of mobility as  
41  
42 17 evidenced when comparing mean TUG and gait speed performance to normative data for  
43  
44  
45 18 community-dwelling adults in Ireland<sup>1</sup>. Norms for RCS are not available for the Irish population,  
46  
47 19 but average performance was slightly slower than age-matched norms presented in the literature  
48  
49  
50 20 <sup>33</sup> although wide variation in testing protocols has been recognised <sup>34</sup>.

### 54 22 *Timed Up-and-go*

57 23 TUG did not vary between baseline and repeat assessments or by time of day, however there was  
58  
59 24 a significant rater effect with a difference of 1.22 s ( $P<.001$ ) between the two nurses. The  
60

1  
2 1 between-person SD was 1.31 s. The SEM was 0.75 s, leading to moderate-good reliability  
3  
4  
5 2 (ICC=0.75) and MDC estimates of 1.75 s at the 90% level and 2.08 s at the 95% level. This means  
6  
7 3 that a difference of 1.75-2.08 s between two assessments in the same individual can be expected  
8  
9  
10 4 by chance depending on the confidence interval used and when controlling for all other factors  
11  
12 5 (rater, time between assessments and time of day). Analysis of TUG on a logarithmic scale  
13  
14 6 suggests similar reliability (ICC=0.71), and a SEM of 0.09. The MDC<sub>95</sub> of 0.24 for log(TUG) suggests  
15  
16  
17 7 that a relative change in TUG of up to 27% (the inverse logarithm of 0.24 is 1.27) might be  
18  
19 8 expected by chance in 95% of paired samples. This finding is applicable across the spectrum of  
20  
21  
22 9 baseline TUG scores.  
23  
24  
25 10

#### 26 11 *Repeated chair stands*

27 12 RCS was completed slightly more quickly at the repeat measurement (difference=0.47 s,  $P=.04$ )  
28  
29 13 and when the assessment was carried out by Nurse 1 (difference=1.09 s,  $P<.001$ ) but did not vary  
30  
31  
32 14 with time of day. The ICC was 0.66 and SEM was 1.63 s while MDC was estimated to be 3.80 s at  
33  
34  
35 15 the 90% level and 4.52 s at the 95% level. Time to complete RCS was also analysed on the log  
36  
37  
38 16 scale, where reliability was similar (ICC=0.68), SEM was 0.13 and MDC was 0.35 at the 95%  
39  
40  
41 17 confidence level.  
42  
43  
44 18

#### 45 19 *Usual gait speed*

46 20 UGS did not vary between baseline and repeat assessment or by time of day, however there was a  
47  
48  
49 21 significant rater effect with a difference of 7.36 cm/s ( $P<.001$ ). Reliability was good (ICC=0.88) as  
50  
51  
52 22 the between-person SD (18.65 cm/s) was much higher than the SEM (7.03 cm/s), resulting in a  
53  
54  
55 23 MDC<sub>90</sub> of 16.40 cm/s and MDC<sub>95</sub> of 19.49 cm/s.  
56  
57  
58 24  
59  
60

### 1 *Manual dual task gait speed*

2 Gait speed became less reliable as the complexity of the dual task conditions increased. MGS was  
3 consistent across repeat assessments but varied by rater (difference=4.88 cm/s,  $P=.02$ ) and time of  
4 day (difference=3.62 s,  $P=0.03$ ). ICC was lower than was observed for UGS (ICC=0.83), SEM was  
5 higher (8.97 cm/s) and consequently so was MDC<sub>90</sub> (20.93 cm/s) and MDC<sub>95</sub> (24.87 cm/s).

### 6 *Cognitive dual task gait speed*

7 CGS did not vary by repeat assessment, rater or time of day, however reliability estimates were  
8 poorest out of all gait speed measures (ICC=0.77; SEM=12.53 cm/s; MDC<sub>95</sub>=34.73 cm/s).

9  
10  
11 For all observed rater effects, including those where performance was automatically measured  
12 (i.e. with GAITRite), participants completed the mobility tasks more quickly when assessed by  
13 Nurse 1.

## 14 **Discussion**

15 We report test-retest reliability, SEM and MDC of commonly used mobility tests in a sample of  
16 relatively healthy, community-dwelling Irish adults aged 50 years and older. We found good test-  
17 retest reliability for walking speed and motor dual task walking speed and moderate-good  
18 reliability for TUG and cognitive dual task walking speed however, the lowest ICC was observed for  
19 RCS. These findings contrast to previous studies which reported moderate to excellent reliability  
20 for all of these measures<sup>9-11 18 19 21-25 35</sup>. As ICC depends on the distribution of scores within the  
21 sample it is estimated in and reflects relative reliability, it is specific to that particular setting and  
22 population<sup>8</sup>. Lower reliability here is likely to reflect more homogeneous population  
23



1  
2  
3 1 representative samples (hence lower between-person standard deviations) compared to clinical  
4  
5 2 samples with varying degrees of impairment.  
6

7 3  
8  
9  
10 4 SEM and MDC provide an indication of absolute reliability. MDC allows the assessor to interpret if  
11  
12 5 an observed change score is above that expected due to measurement error and therefore if it  
13  
14 6 represents a genuine change in performance. In this study, MDC for TUG (2.08 s at the 95% level)  
15  
16  
17 7 is lower than that presented in previous studies of healthy ( $MDC_{95}=4.71$  s)<sup>16</sup> and cognitively  
18  
19 8 impaired ( $MDC_{95}=5.88-6.87$  s) older adults<sup>14 15</sup> and Parkinson's disease patients ( $MDC_{95}=11$  s)<sup>17</sup>.  
20  
21  
22 9 However, reporting variability in TUG as a percentage change in performance rather than in  
23  
24 10 absolute terms may be more appropriate. In contrast,  $MDC_{95}$  for UGS, MGS and CGS  
25  
26  
27 11 ( $MDC_{95}=19.49-34.76$  cm/s) are generally higher than the values estimated in community-dwelling  
28  
29 12 healthy adults ( $MDC_{95}=13.6$  cm/s)<sup>22</sup>, community-dwelling and hospitalised fallers ( $MDC_{95}=12.4-$   
30  
31 13  $15.5$  cm/s)<sup>35</sup> and in those post-stroke ( $MDC_{95}=20$  cm/s)<sup>36</sup>. These differences may be due to the  
32  
33  
34 14 position on the performance scale as participants in these studies demonstrated poorer mobility  
35  
36  
37 15 than participants in the SHARE-TILDA study<sup>22 35 36</sup>.  
38

39 16  
40  
41 17 Many longitudinal or intervention based studies vary widely in sample characteristics, co-  
42  
43  
44 18 morbidity and time intervals between assessments. This makes cross-study comparisons difficult  
45  
46 19 and therefore reliability measures are best estimated for each sample and for groups with specific  
47  
48  
49 20 diagnoses. This study provides guidance on MDC across the range of function in a generally  
50  
51 21 healthy, population-based sample, when measurements are compared weeks or months apart.  
52  
53  
54 22 These estimates should be used when assessing individual changes in mobility performance over  
55  
56 23 this time-scale e.g. when examining the effects of an intervention or patient progression, when  
57  
58  
59 24 calculating required sample sizes for studies using these outcomes or when applying methods to  
60

1  
2  
3 1 adjust for measurement error in epidemiological studies. Participants in this study were relatively  
4  
5 2 healthy and while acute changes in health and performance can occur even with shorter follow-  
6  
7 3 up, they are unlikely to demonstrate a consistent, genuine change in performance in the time  
8  
9  
10 4 period examined. While using a shorter time period and/or same-day repeated measurements  
11  
12 5 would likely provide higher estimates of reliability, this approach was taken to reflect the variation  
13  
14 6 that is likely to be observed in real-world clinical and research settings over a longer time period.  
15  
16

17 7  
18  
19 8 These results show the significant effect of inter-rater variation even with two highly trained and  
20  
21  
22 9 experienced research nurses. This suggests that changing rater introduces additional variance in  
23  
24 10 the measures beyond within-participant variation. The effect was observed in the GAITRite®  
25  
26  
27 11 assessment as well as stopwatch based tests suggesting that rater differences in reaction time do  
28  
29 12 not explain this. Both nurses were highly experienced and followed standardised protocols,  
30  
31  
32 13 however one explanation could be that they have different styles of interaction with respondents,  
33  
34 14 which may have impacted on the respondent's understanding of the task, or their motivation and  
35  
36  
37 15 subsequent desire to perform well. This emphasises the importance of providing appropriate  
38  
39 16 training for all raters to ensure that measurements are as accurate and consistent as possible. In  
40  
41 17 an effort to detect and address these differences, studies could examine within-day rater  
42  
43  
44 18 differences on a small number of participants although only a limited number of tests would be  
45  
46  
47 19 feasible to avoid fatigue effects. Where possible, analyses should also be adjusted to account for  
48  
49 20 differences between the raters conducting the assessments.  
50

51 21

## 52 53 22 ***Study Strengths and Limitations***

54  
55  
56 23 A strength of this study is the population-based sample of relatively healthy middle-aged and older  
57  
58  
59 24 adults used in the analysis. In addition, our estimates of reliability remove time of day and rater  
60

1  
2  
3 1 effects. For measures that are skewed, a different MDC may be required depending on whether  
4  
5 2 performance is at the higher or lower ends of the spectrum. To account for this, we represent  
6  
7 3 relevant findings on the multiplicative (logarithmic) scale and the additive scale. Although a  
8  
9 4 stopwatch is the easiest and most cost effective way to measure gait speed, the GAITRite® mat is  
10  
11 5 frequently used in research. Therefore, this analysis provides useful guidance on data obtained  
12  
13 6 using simple and more complex instruments. However, there are also a number of limitations in  
14  
15 7 this study. Participants were not asked to restrict their exercise levels, activities or medications  
16  
17 8 before the assessments, all of which could contribute to measurement variation. While the  
18  
19 9 participants did not report any injuries that prevented them from doing the tests, it is also possible  
20  
21 10 that they may have had a low level injury or have been recovering from an injury at either  
22  
23 11 assessment which may account for some of the within-subject variation observed.  
24  
25  
26  
27  
28  
29  
30  
31

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

24

## Conclusion

Gait speed obtained during normal walking conditions and when completing a manual dual task are repeatable when performed at time intervals of several weeks to months, with lower reliability observed for the cognitive dual walk, TUG and RCS. There is also a potentially large effect of rater, even for measures that are automatically measured. The estimates of MDC are presented for a population based sample of relatively healthy middle-aged and older Irish adults and can be used to assess changes in performance in individuals drawn from comparable populations. Similar robust reliability studies are recommended to inform the use and interpretation of repeated assessments in other populations such as those with specific co-morbidities. Additional analysis using anchor-based approaches could be used to examine if these changes are of clinical importance.

## 1 Author contributions:

2 Substantial contributions to the conception or design of the work; or the acquisition, analysis, or  
3 interpretation of data for the work – OD, GS, AB-S, RAK; Drafting the work or revising it critically  
4 for important intellectual content – OD, GS, AB-S, RAK; Final approval of the version to be  
5 published – OD, GS, AB-S, RAK; Agreement to be accountable for all aspects of the work in  
6 ensuring that questions related to the accuracy or integrity of any part of the work are  
7 appropriately investigated and resolved – OD, GS, AB-S, RAK.

## 10 References

- 11 1. Kenny RA, Coen RF, Frewen J, et al. Normative values of cognitive and physical function in older  
12 adults: findings from the Irish Longitudinal Study on Ageing. *Journal of the American*  
13 *Geriatric Society* 2013;61(2):S279-S90.
- 14 2. Abellan Van Kan G, Rolland Y, Andrieu S, et al. Gait speed at usual pace as a predictor of adverse  
15 outcomes in community-dwelling older people an International Academy on Nutrition and  
16 Aging (IANA) Task Force. *Journal of Nutrition Health and Aging* 2009;13(10):881-89. doi:  
17 10.1007/s12603-009-0246-z
- 18 3. Cooper R, Kuh D, Hardy R, et al. Objectively measured physical capability levels and mortality:  
19 systematic review and meta-analysis. *BMJ* 2010;341(341):c4467. doi: 10.1136/bmj.c4467
- 20 4. Cooper R, Kuh D, Cooper C, et al. Objective measures of physical capability and subsequent  
21 health: a systematic review. *Age Ageing* 2011;40(1):14-23.
- 22 5. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between  
23 reproducibility and responsiveness. *Qual Life Res* 2001;10(7):571-8.
- 24 6. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30(1):1-  
25 15.

- 1  
2  
3 1 7. Glymour MM, Weuve J, Berkman LF, et al. When is baseline adjustment useful in analyses of  
4  
5 2 change? An example with education and cognitive change. *Am J Epidemiol*  
6  
7 3 2005;162(3):267-78.
- 8  
9  
10 4 8. Rydwik E, Bergland A, Forsen L, et al. Investigation into the reliability and validity of the  
11  
12 5 measurement of elderly people's clinical walking speed: a systematic review. *Physiother*  
13  
14 6 *Theory Pract* 2012;28(3):238-56.
- 15  
16  
17 7 9. Griswold D, Rockwell K, Killa C, et al. Establishing the reliability and concurrent validity of  
18  
19 8 physical performance tests using virtual reality equipment for community-dwelling healthy  
20  
21 9 elders. *Disability and Rehabilitation* 2014;25:1-5. doi: doi:10.3109/09638288.2014.952451
- 22  
23  
24 10 10. Regterschot GRH, Zhang W, Baldus H, et al. Test-retest reliability of sensor-based sit-to-stand  
25  
26 11 measures in young and older adults. *Gait Posture* 2014;40(1):220-24. doi:  
27  
28 12 <http://dx.doi.org/10.1016/j.gaitpost.2014.03.193>
- 29  
30  
31 13 11. Ng SS, Hui-Chan CW. The Timed Up & Go Test: Its Reliability and Association With Lower-Limb  
32  
33 14 Impairments and Locomotor Capacities in People With Chronic Stroke. *Arch Phys Med*  
34  
35 15 *Rehabil* 2005;86(8):1641-47. doi: 10.1016/j.apmr.2005.01.011
- 36  
37  
38  
39 16 12. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail  
40  
41 17 elderly persons. *J Am Geriatr Soc* 1991;39(2):142-48. [published Online First: PMID:  
42  
43 18 1991946 ]
- 44  
45  
46 19 13. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-  
47  
48 20 Dwelling Older Adults Using the Timed Up & Go Test. *Phys Ther* 2000;80(9):896-903.
- 49  
50  
51 21 14. Blankevoort CG, van Heuvelen MJ, Scherder EJ. Reliability of six physical performance tests in  
52  
53 22 older people with dementia. *Phys Ther* 2013;93(1):69-78.

- 1  
2  
3 1 15. Ries JD, Echternach JL, Nof L, et al. Test-retest reliability and minimal detectable change scores  
4  
5 2 for the timed "up & go" test, the six-minute walk test, and gait speed in people with  
6  
7 3 Alzheimer disease. *Phys Ther* 2009;89(6):569-79.  
8  
9  
10 4 16. Mangione KK, Craik RL, McCormick AA, et al. Detectable Changes in Physical Performance  
11  
12 5 Measures in Elderly African Americans. *Phys Ther* 2010;90(6):921-27. doi:  
13  
14 6 10.2522/ptj.20090363  
15  
16  
17 7 17. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and  
18  
19 8 ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease  
20  
21 9 rating scale in people with parkinsonism. *Phys Ther* 2008;88(6):733-46.  
22  
23  
24 10 18. Goldberg A, Chavis M, Watkins J, et al. The five-times-sit-to-stand test: validity, reliability and  
25  
26 11 detectable change in older females. *Aging Clin Exp Res* 2012;24(4):339-44.  
27  
28  
29 12 19. Hollman JH, Childs KB, McNeil ML, et al. Number of strides required for reliable measurements  
30  
31 13 of pace, rhythm and variability parameters of gait during normal and dual task walking in  
32  
33 14 older individuals. *Gait Posture* 2010;32(1):23-28. doi: 10.1016/j.gaitpost.2010.02.017  
34  
35  
36 15 20. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
37  
38 16 in community-dwelling and hospitalized older fallers. *Gait Posture* (38(4):1010-4)  
39  
40  
41 17 21. Brach JS, Perera S, Studenski S, et al. The Reliability and Validity of Measures of Gait Variability  
42  
43 18 in Community-Dwelling Older Adults. *Arch Phys Med Rehabil* 2008;89(12):2293-96. doi:  
44  
45 19 10.1016/j.apmr.2008.06.010  
46  
47  
48 20 22. Goldberg A, Schepens S. Measurement error and minimum detectable change in 4-meter gait  
49  
50 21 speed in older adults. *Aging Clin Exp Res* 2011;23(5-6):406-12.  
51  
52  
53 22 23. Menz HB, Lord SR, St George R, et al. Walking stability and sensorimotor function in older  
54  
55 23 people with diabetic peripheral neuropathy. *Arch Phys Med Rehabil* 2004;85(2):245-52.  
56  
57 24 doi: 10.1016/j.apmr.2003.06.015  
58  
59  
60

- 1  
2  
3 1 24. Van Iersel MB, Benraad CEM, Olde Rikkert MGM. Validity and reliability of quantitative gait  
4  
5 2 analysis in geriatric patients with and without dementia. . *J Am Geriatr Soc*  
6  
7 3 2007;55(4):632-34. doi: 10.1111/j.1532-5415.2007.01130.x  
8  
9  
10 4 25. Paterson KL, Hill KD, Lythgo ND, et al. The reliability of spatiotemporal gait data for young and  
11  
12 5 older women during continuous overground walking. *Arch Phys Med Rehabil*  
13  
14 6 2008;89(12):2360-5.  
15  
16  
17 7 26. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: The Survey of Health,  
18  
19 8 Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*  
20  
21 9 2013;42(4):992-1001.  
22  
23  
24 10 27. UCD Geary Institute and Irish Centre for Social Gerontology N. SHARE Ireland - Survey of  
25  
26 11 Health, Ageing and Retirement in Europe [Internet]. Available from:  
27  
28 12 [http://gearyucdie/share/fileadmin/user\\_upload/shareresults/Share\\_Wave1\\_Resultspdf](http://gearyucdie/share/fileadmin/user_upload/shareresults/Share_Wave1_Resultspdf)  
29  
30 13 2008  
31  
32  
33  
34 14 28. Feeney J, Savva GM, O'Regan C, et al. Measurement Error, Reliability, and Minimum  
35  
36 15 Detectable Change in the Mini-Mental State Examination, Montreal Cognitive Assessment,  
37  
38 16 and Color Trails Test among Community Living Middle-Aged and Older Adults. *J Alzheimers*  
39  
40 17 *Dis* 2016;53(3):1107-14. doi: 10.3233/jad-160248 [published Online First: 2016/06/04]  
41  
42  
43  
44 18 29. Finucane C, Savva GM, Kenny RA. Reliability of orthostatic beat-to-beat blood pressure tests:  
45  
46 19 implications for population and clinical studies. *Clin Auton Res* 2017;27(1):31-39. doi:  
47  
48 20 10.1007/s10286-016-0393-3 [published Online First: 2017/01/14]  
49  
50  
51 21 30. Cronin H, O'Regan C, Finucane C, et al. Health and aging: development of the Irish Longitudinal  
52  
53 22 Study on Ageing health assessment. *J Am Geriatr Soc* 2013;61(2):12197.  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 1 31. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing  
4  
5 2 lower extremity function: association with self-reported disability and prediction of  
6  
7 3 mortality and nursing home admission. *J Gerontol* 1994;49(2):M85-94.  
8  
9  
10 4 32. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for  
11  
12 5 Reliability Research. *J Chiropr Med* 2016;15(2):155-63. doi: 10.1016/j.jcm.2016.02.012  
13  
14 6 [published Online First: 2016/03/31]  
15  
16  
17 7 33. Bohannon RW. Reference values for the five-repetition sit-to-stand test: a descriptive meta-  
18  
19 8 analysis of data from elders. *Perceptual and motor skills* 2006;103(1):215-22. doi:  
20  
21 9 10.2466/pms.103.1.215-222 [published Online First: 2006/10/14]  
22  
23  
24 10 34. Mehmet H, Yang AWH, Robinson SR. What is the optimal chair stand test protocol for older  
25  
26 11 adults? A systematic review. *Disability and Rehabilitation* 2019:1-8. doi:  
27  
28 12 10.1080/09638288.2019.1575922  
29  
30  
31 13 35. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
32  
33 14 in community-dwelling and hospitalized older fallers. *Gait Posture* 2013;38(4):1010-4.  
34  
35  
36 15 36. Lewek MD, Randall EP. Reliability of spatiotemporal asymmetry during overground walking for  
37  
38 16 individuals following chronic stroke. *J Neurol Phys Ther* 2011;35(3):116-21.  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Table 1. Mobility performance scores obtained at baseline and repeat assessments, with different raters and at different times of day.

	Assessment		Rater <sup>a</sup>		Time of day <sup>b</sup>	
	Baseline	Repeat	Nurse 1	Nurse 2	Test AM	Test PM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
TUG (s)	8.88 (1.39)	8.87 (1.54)	8.13 (1.20)	9.35 (1.51)***	8.83 (1.49)	8.69 (1.25)
log(TUG)	2.17 (0.02)	2.17 (0.01)	2.08 (0.02)	2.22 (0.02)***	2.16 (0.02)	2.15 (0.02)
RCS (s)	12.49 (2.87)	12.02 (2.48)*	11.80 (2.27)	12.89 (2.88)***	12.17 (2.99)	12.00 (2.46)
logRCS	2.50 (0.22)	2.46 (0.21)*	2.45 (0.20)	2.53 (0.24)**	2.47 (0.24)	2.46 (0.22)
UGS (cm/s)	137.95 (20.21)	138.20 (19.32)	145.82 (18.94)	138.46 (17.85)***	137.62 (17.68)	137.74 (17.38)
MGS (cm/s)	116.76 (21.84)	118.71 (19.93)	123.07 (18.95)	118.07 (20.45)**	117.86 (19.85)	122.19 (17.21)*
CGS (cm/s)	115.23 (24.08)	115.15 (25.21)	118.29 (25.24)	117.40 (20.99)	117.45 (24.01)	118.84 (20.18)

Notes: TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task gait speed

<sup>a</sup> Rater scores are calculated only among participants who changed rater at the repeat assessment

<sup>b</sup> Time of day scores are calculated only among participants who changed time of day at the repeat assessment.

\*  $P < 0.05$ ; \*\*  $P < 0.01$ ; \*\*\*  $P < 0.001$

Table 2. Variance and reliability estimates for all mobility tests.

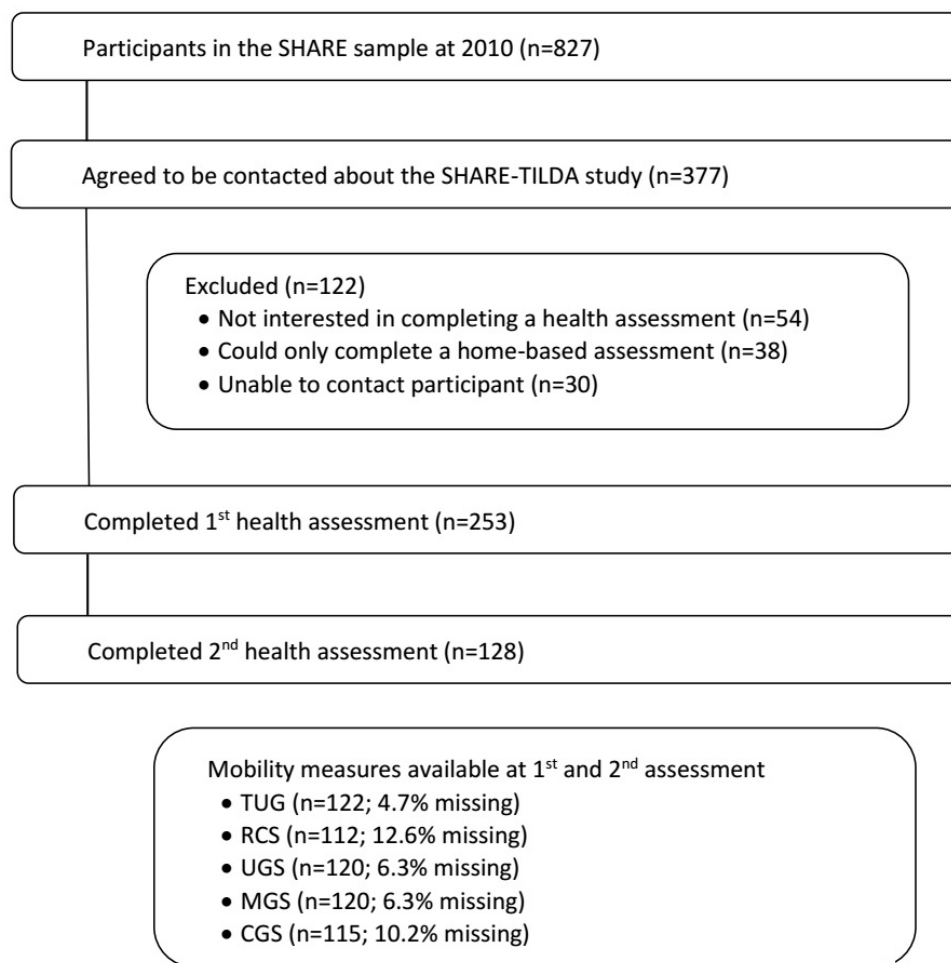
	<b>SD<sub>between</sub> (95% CI)</b>	<b>SEM (95% CI)</b>	<b>ICC (95% CI)</b>	<b>MDC<sub>90</sub></b>	<b>MDC<sub>95</sub></b>
TUG (s)	1.31 (1.12-1.52)	0.75 (0.66-0.85)	0.75 (0.66-0.82)	1.75	2.08
logTUG	0.13 (0.11-0.15)	0.09 (0.08-0.10)	0.71 (0.61-0.79)	0.2	0.24
RCS (s)	2.29 (1.93-2.70)	1.63 (1.43-1.86)	0.66 (0.55-0.76)	3.8	4.52
logRCS	0.18 (0.16-0.22)	0.13 (0.11-0.14)	0.68 (0.57-0.77)	0.29	0.35
UGS (cm/s)	18.65 (16.34-21.29)	7.03 (6.20-7.98)	0.88 (0.83-0.91)	16.4	19.49
MGS (cm/s)	19.57 (17.04-22.46)	8.97 (7.90-10.19)	0.83 (0.76-0.88)	20.93	24.87
CGS (cm/s)	22.73 (19.62-26.34)	12.53 (10.99-14.28)	0.77 (0.68-0.83)	29.24	34.73

Notes: SEM, standard error of the measurement; TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task speed; ICC, intra-class correlation; MDC, minimum detectable change

1  
2 Figure 1: Exclusion criteria used to establish eligible participants for this analysis.  
3

4  
5 Note: CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair  
6  
7 stands; SHARE, Survey for Health Ageing and Retirement in Europe; TILDA, The Irish Longitudinal  
8  
9 Study on Ageing; TUG, Timed Up-and-Go; UGS, usual gait speed.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only



Exclusion criteria used to establish eligible participants for this analysis.

245x247mm (96 x 96 DPI)

1 STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation	Location
<b>Title and abstract</b>	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	P1, P2
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	P2
<b>Introduction</b>			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	P5-7
Objectives	3	State specific objectives, including any prespecified hypotheses	P6
<b>Methods</b>			
Study design	4	Present key elements of study design early in the paper	P7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	P7-8
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	P7-8
		(b) For matched studies, give matching criteria and number of exposed and unexposed	-
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	P8-10
Data sources/measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	P8-10
Bias	9	Describe any efforts to address potential sources of bias	P10
Study size	10	Explain how the study size was arrived at	P7-8
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	P10-11
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	P10-11
		(b) Describe any methods used to examine subgroups and interactions	P10-11
		(c) Explain how missing data were addressed	P10
		(d) If applicable, explain how loss to follow-up was addressed	-
		(e) Describe any sensitivity analyses	-
<b>Results</b>			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	P8, 12, Fig 1
		(b) Give reasons for non-participation at each stage	P8
		(c) Consider use of a flow diagram	Fig 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	P12
		(b) Indicate number of participants with missing data for each variable of interest	Fig 1
		(c) Summarise follow-up time (eg, average and total amount)	P12
Outcome data	15*	Report numbers of outcome events or summary measures over time	P23
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear	P12, 14, 24

		which confounders were adjusted for and why they were included	
		(b) Report category boundaries when continuous variables were categorized	-
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	-
<b>Discussion</b>			
Key results	18	Summarise key results with reference to study objectives	P12-14
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	P17
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	P14-17
Generalisability	21	Discuss the generalisability (external validity) of the study results	P15-17
<b>Other information</b>			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	P4

\*Give information separately for exposed and unexposed groups.

**Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at <http://www.strobe-statement.org>.

# BMJ Open

## Reliability, measurement error and minimum detectable change in mobility measures: a cohort study of community-dwelling adults aged 50 years and over in Ireland

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-030475.R2
Article Type:	Original research
Date Submitted by the Author:	04-Oct-2019
Complete List of Authors:	Donoghue, Orna; University of Dublin Trinity College, The Irish Longitudinal Study on Ageing (TILDA) Savva, George; Quadram Institute Bioscience, Börsch-Supan, Axel; Max Planck Institute for Social Law and Social Policy, Munich Center for the Economics of Aging Kenny, RoseAnne; Trinity College Dublin, The Irish Longitudinal Study on Ageing (TILDA); St James Hospital, Mercer's Institute for Successful Ageing
<b>Primary Subject Heading</b>:	Geriatric medicine
Secondary Subject Heading:	Geriatric medicine
Keywords:	repeatability, physical performance tests, longitudinal change, Epidemiology < TROPICAL MEDICINE

SCHOLARONE™  
Manuscripts

1  
2  
3 1 **Reliability, measurement error and minimum detectable change in mobility measures: a cohort**  
4  
5 2 **study of community-dwelling adults aged 50 years and over in Ireland**  
6

7 3  
8  
9  
10 4 Orna A Donoghue, PhD <sup>a</sup>, George M Savva, PhD <sup>b</sup>, Axel Börsch-Supan, PhD <sup>c</sup>, Rose Anne Kenny, MD  
11  
12 5 <sup>a,d</sup>

13  
14 6  
15  
16  
17 7 **Affiliations:**

18  
19 8 <sup>a</sup> The Irish Longitudinal Study on Ageing (TILDA), Trinity College Dublin, Lincoln Place, Dublin 2,  
20  
21  
22 9 Ireland. Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie).

23  
24 10 <sup>b</sup> Quadram Institute Bioscience, Norwich Research Park, Norwich, UK. Email:  
25  
26  
27 11 [George.savva@quadram.ac.uk](mailto:George.savva@quadram.ac.uk).

28  
29 12 <sup>c</sup> Munich Center for the Economics of Aging, Max-Planck Institute for Social Law and Social Policy,  
30  
31  
32 13 Amalienst, Munich, Germany. Email: [axel@boersch-supan.de](mailto:axel@boersch-supan.de).

33  
34 14 <sup>a,d</sup> Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland. Email:  
35  
36 15 [rkenny@tcd.ie](mailto:rkenny@tcd.ie).

37  
38 16  
39  
40 17 **Corresponding author:** Dr Orna Donoghue, The Irish Longitudinal Study on Ageing (TILDA), Trinity  
41  
42  
43 18 College Dublin, Lincoln Place, Dublin 2, Ireland  
44  
45 19 Tel: +353 1 896 4391; Fax: +353 1 896 2451; Email: [odonogh@tcd.ie](mailto:odonogh@tcd.ie)  
46  
47

48 20  
49  
50 21 **Acknowledgements:** The authors would like to acknowledge the contribution of the participants  
51  
52  
53 22 and members of the TILDA and SHARE teams.  
54

55 23  
56  
57 24 **Word count:** 3161  
58  
59  
60 25



## 1 Abstract

2 Objective: To estimate the effects of repeat assessments, rater and time of day on mobility  
3 measures and to estimate their variation between- and within- participants in a population-based  
4 sample of Irish adults aged  $\geq 50$  years.

5 Design: Test-retest study in a population representative sample.

6 Setting: Academic health assessment centre of The Irish Longitudinal Study on Ageing (TILDA).

7 Participants: 128 community-dwelling adults from the Survey for Health, Ageing and Retirement in  
8 Europe (SHARE) Ireland study who agreed to take part in the SHARE-Ireland / TILDA collaboration.

9 Interventions: Not applicable.

10 Outcome Measures: Participants performed Timed Up-and-Go (TUG), repeated chair stands (RCS)  
11 and walking speed tests administered by one of two raters. Repeat assessments were conducted  
12 1-4 months later. Participants were randomised with respect to a change in time (morning,  
13 afternoon) and whether the rater was changed between assessments. Within- and between-  
14 participant variance for each measure was estimated using mixed effects models. Intra-class  
15 correlation (ICC), standard error of measurement (SEM) and minimum detectable change (MDC)  
16 were reported.

17 Results: Average performance did not vary between baseline and repeat assessments in any test,  
18 except RCS. The rater significantly affected performance on all tests except one, but time of day  
19 did not. Reliability varied from ICC=0.66 (RCS) to ICC=0.88 (usual gait speed). MDC was 2.08 s for  
20 TUG, 4.52 s for RCS and ranged from 19.49-34.73 cm/s for walking speed tests. There was no  
21 evidence for lower reliability of gait parameters with increasing time between assessments.

22 Conclusions: Reliability varied for each test when measurements are obtained over 1-4 months  
23 with most variation due to rater effects. Usual and motor dual task gait speed demonstrated  
24 highest reliability.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1 **Key words:** repeatability, physical performance tests, longitudinal change, epidemiology

2

For peer review only

## 1 Article summary

### 2 Strengths and limitations of this study

- 3 • This study provides information on the effects of repeat assessments, rater and time of day  
4 on test-retest reliability of mobility measures obtained over 1-4 months using a population-  
5 based sample of relatively healthy middle-aged and older adults aged  $\geq 50$  years in Ireland.
- 6 • The use of common tests such as Timed Up-and-Go, repeated chair stands and GAITRite  
7 assessments makes this analysis relevant for other studies looking at change in mobility.
- 8 • Mixed effects models were used to estimate within- and between-participant variance for  
9 each measure allowing intra-class correlation (ICC) and standard error of measurement  
10 (SEM) and minimum detectable change (MDC) to be presented, net of fixed effects.
- 11 • For some measures, MDC was presented on the multiplicative (logarithmic) scale as well as  
12 the natural additive scale to account for skewness and to ensure that findings are  
13 applicable across all levels of performance.
- 14 • Changes in exercise levels, activities, medications and current injury status could have  
15 contributed to measurement variation but these were not measured. However the fact  
16 that the measures did not become less reliable with increasing time since assessments  
17 suggests that this does not substantially affect the findings.

1  
2  
3 1 **Funding:** TILDA received financial support from the Irish Government (Department of Health and  
4  
5 2 Children), the Atlantic Philanthropies and Irish Life plc. The SHARE-TILDA project was funded by  
6  
7 3 the National Institute of Aging (Prime Award Number R21AG040387). Funders had no involvement  
8  
9  
10 4 in analysis and preparation of this paper.  
11  
12 5

13  
14 6 **Competing interests:** None declared  
15  
16  
17 7

18  
19 8 **Data sharing statement:** The anonymised SHARE-TILDA dataset is available through the on-site  
20  
21  
22 9 “hot desk” facility at TILDA, Trinity College Dublin. Researchers should contact [tilda@tcd.ie](mailto:tilda@tcd.ie) for  
23  
24 10 more information.  
25  
26  
27 11  
28  
29 12  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 1 Introduction

2 Performance based measures such as Timed Up-and-Go (TUG), repeated chair stands (RCS) and  
3 walking speed tests are commonly used to assess mobility and lower limb function of older adults  
4 in clinical and research settings<sup>1</sup>. These measures are good predictors of falls, disability, cognitive  
5 decline and mortality<sup>2-4</sup>. To be useful, they also need to be reliable (consistent when measured on  
6 several occasions and when there is no change in a subject's underlying performance) and  
7 responsive (able to detect a change when there is one)<sup>5</sup>. Good reliability allows changes in  
8 measurements to be tracked over time<sup>6</sup>.

9  
10 However, all tests are subject to measurement error due to within-subject, inter-trial, inter-rater  
11 effects. They are also liable to day-to-day variation due to patient level factors that do not reflect  
12 the underlying risk status that they are attempting to measure. This has several implications.  
13 Clinically, if an individual improves or declines between two testing sessions, it is important to  
14 know how likely it is that the observed change is a genuine change in status and is not due to  
15 measurement error or a transient effect. In research settings, unreliable measures can lead to  
16 regression dilution bias or false positive associations when testing predictors of longitudinal  
17 change<sup>7</sup>. To account for this, several measures of relative reliability i.e. intra-class correlation  
18 (ICC), and absolute reliability i.e. standard error of measurement (SEM) and minimum detectable  
19 change (MDC), are often reported<sup>8</sup>.

20  
21 SEM is the standard deviation of the measurement error of a measure within an individual, for a  
22 given 'true' value of the underlying construct. The SEM determines the MDC, which is the smallest  
23 difference between two single observations that can be confidently attributed to a genuine  
24 difference and not to measurement error. ICC is a measure of the proportion of variance within a

1  
2 1 population that is attributable to variance across individuals as opposed to measurement error  
3  
4  
5 2 within individuals. As opposed to SEM and MDC, ICC depends on both the SEM and the variation  
6  
7 3 between members of a sample, and so is not usually comparable or applicable across samples with  
8  
9  
10 4 different levels of heterogeneity.  
11

12 5  
13  
14 6 The within-session and one week test-retest reliability of TUG in community-dwelling, older adults  
15  
16  
17 7 is well known, and is known to be high ( $ICC \geq 0.96$ )<sup>9-11</sup> in various populations as is the inter-rater<sup>12</sup>  
18  
19 8<sup>13</sup> and intra-rater reliability<sup>12</sup>. MDC at the 95% confidence level ( $MDC_{95}$ ) has been reported to vary  
20  
21  
22 9 between 3.33-6.87 s in healthy and cognitively impaired older adults<sup>14-16</sup> and up to 11 sec in  
23  
24 10 Parkinson's disease patients<sup>17</sup>. The within-session test-retest reliability of RCS is also very high  
25  
26  
27 11 ( $ICC=0.93-0.95$ )<sup>9,18</sup>, however SEM and MDC for community-dwelling adults are not known.  
28

29 12  
30  
31  
32 13 Walking speed can be measured using stopwatches, timing gates or sensed mats. The test-retest  
33  
34 14 reliability of usual gait speed (UGS) measured using a GAITRite® walkway has been reported to be  
35  
36  
37 15 between  $ICC=0.84$  and  $0.97$  for assessments given up to two weeks apart<sup>19-25</sup>. Similar values have  
38  
39 16 been reported for one hour test-retest reliability of dual task gait speed ( $ICC=0.85-0.93$ )<sup>19,20</sup>.

40  
41 17 Fewer studies have reported SEM or MDC in healthy populations with MDC values of 12.4-13.6  
42  
43  
44 18 cm/s reported for UGS<sup>20,22</sup> and 15.5 cm/s for dual task gait speed<sup>20</sup>. However, reliability of dual  
45  
46 19 task gait speed may also be dependent on the actual dual task and therefore is not comparable  
47  
48  
49 20 across studies unless the same test has been used.  
50

51 21  
52  
53  
54 22 Here we report the test-retest reliability measured by ICC, SEM and MDC in a pragmatic  
55  
56 23 epidemiologic setting. We explore how reliability changes when lag between assessments varies  
57  
58  
59 24 between 1 month and 4 months, when rater changes or is held constant, and whether or not time  
60

1  
2  
3 1 of assessment varies, in a large sample of healthy adults aged 50 and older recruited at random  
4  
5 2 from the population.  
6

7 3  
8  
9  
10 4 In epidemiologic settings, these measures are commonly used as proxies for the underlying  
11  
12 5 general cognitive and physical health status of participants around the time of the assessment.  
13  
14 6 Short-term fluctuations in these measures, for example due to acute illness or day-to-day  
15  
16  
17 7 variation, add error to these outcomes along with measurement error associated with the  
18  
19 8 instruments themselves. Hence when comparing measures over longer time periods, that is, years  
20  
21  
22 9 or decades typical of epidemiologic research, it is important to know how well single measures of  
23  
24 10 physical and cognitive function reflect the underlying health status of the participant, net of any  
25  
26  
27 11 factors that might cause a short-term fluctuation. Therefore, we tested the concordance between  
28  
29 12 pairs of measures between one and four months apart, to estimate the error association with both  
30  
31  
32 13 measurement and day-to-day fluctuation in each measure. Understanding natural variation in  
33  
34 14 outcomes over one to four months is also essential when planning clinical trials with follow-up  
35  
36  
37 15 time in this range, since this is the natural variation against which any treatment effect would be  
38  
39 16 compared.  
40

## 41 17 42 43 44 18 **Methods**

### 45 46 19 ***Participants***

47  
48  
49 20 Participants were a subsample from the Survey of Health, Ageing and Retirement in Europe  
50  
51 21 (SHARE), a longitudinal, cross-national study on health, socio-economic status and social and  
52  
53  
54 22 family networks of more than 80,000 individuals aged 50 years and over across Europe<sup>26</sup>. The  
55  
56 23 SHARE-Ireland sample (n=1,119) was recruited in Ireland between 2006 and 2007 with a response  
57  
58  
59 24 rate of 55%<sup>27</sup>. A collaboration between SHARE-Ireland and The Irish Longitudinal Study on Ageing  
60

1  
2 1 (TILDA) was established to understand the measurement properties of a comprehensive health  
3  
4  
5 2 assessment among a representative sample of the European population. Reliability of cognitive  
6  
7 3 measures and blood pressure dynamics based on this sample have been published previously<sup>28 29</sup>.  
8  
9

10 4  
11  
12 5 The extant SHARE-Ireland cohort at 2010 (n=827) was contacted and invited to take part in a  
13  
14 6 health assessment that included the same tests and followed the same protocols as those used by  
15  
16  
17 7 TILDA. The health assessment was delivered to the SHARE-Ireland participants by TILDA research  
18  
19 8 nurses within the TILDA health assessment centre based at Trinity College Dublin. Initial contact  
20  
21  
22 9 was made by post and followed up by telephone between September 2011 and March 2012, with  
23  
24 10 377 participants consenting to receive further information about the study. Of these, 253 agreed  
25  
26  
27 11 to an initial health assessment (see Figure 1). Ethical approval for this sub-study was obtained  
28  
29 12 from the Faculty of Health Sciences Research Ethics Committee at Trinity College Dublin. All  
30  
31  
32 13 participants provided informed consent.  
33  
34 14

### 35 36 15 **Health assessments and interview**

37  
38  
39 16 The full health assessment included a 3 hour battery of tests assessing cognitive function, gait and  
40  
41 17 mobility, cardiovascular function and vision<sup>30</sup>. Health assessments were conducted by two highly  
42  
43  
44 18 trained research nurses with approximately 3 years' experience delivering these specific tests in  
45  
46 19 the current setting. Training took approximately one month and nurses used detailed and  
47  
48  
49 20 standardised health assessment protocols which included clear explanations and demonstrations  
50  
51 21 to ensure consistent instructions were provided to all participants. Nurses also underwent periodic  
52  
53  
54 22 quality control procedures to ensure adherence to the protocols.  
55  
56 23



1  
2  
3 1 A short interview was administered by the nurses before the health assessment to capture  
4  
5 2 information on health, chronic disease, disability, employment, social and financial circumstances.  
6  
7 3 Co-morbidity was assessed by asking participants if a doctor had ever told them that they had any  
8  
9  
10 4 of the following conditions: heart attack, high blood pressure, high cholesterol, stroke, diabetes,  
11  
12 5 chronic lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson's disease, cataracts,  
13  
14 6 age related macular degeneration, Alzheimer's disease and atrial fibrillation. The number of  
15  
16  
17 7 conditions was summed and categorised according to 0, 1, 2 or  $\geq 3$  conditions. Participants self-  
18  
19  
20 8 rated their health as excellent, very good, good, fair or poor.  
21  
22 9

23  
24 10 On completing the health assessment, 180 participants were invited to take part in an identical  
25  
26  
27 11 repeat assessment, scheduled after 1-4 months. In total, 128 participants (58 men) agreed to the  
28  
29  
30 12 repeat assessment giving a response rate of 71% (25 refused and 27 were unavailable to attend  
31  
32 13 the repeat assessment within the required timeframe).  
33

34 14  
35  
36 15 Repeat assessments were arranged to distinguish within-person variation from variation caused by  
37  
38  
39 16 changing rater or time of day. The same research nurse conducted the baseline and repeat  
40  
41  
42 17 assessments for half of the participants while another nurse conducted the repeat assessment for  
43  
44 18 the other half of the participants. Time of day when the assessment took place (morning or  
45  
46 19 afternoon) was also changed for half of the participants. Change of rater, change of time of day  
47  
48  
49 20 and delay between assessments (dichotomised at the median) were randomised using a  
50  
51 21 minimisation routine designed to achieve balance across these covariates, as well as the age group  
52  
53  
54 22 and sex of participants. Other factors that could influence performance e.g. health assessment  
55  
56 23 protocols, assessment location, equipment, etc. were held constant across both assessments.  
57  
58  
59 24  
60

### 1 **Physical performance tests**

2  
3  
4  
5 2 Participants completed several mobility tests - TUG, RCS and gait assessments in single and dual  
6  
7 3 task conditions. TUG, which is a common functional mobility test <sup>12</sup> was completed once using  
8  
9  
10 4 walking aids if required. The time taken to rise from the chair (seat height 46 cm), walk 3 m at  
11  
12 5 normal pace, turn around, walk back to the chair and sit down again was recorded using a  
13  
14 6 stopwatch. RCS is an indicator of mobility and lower limb muscular endurance <sup>31</sup>. Participants  
15  
16  
17 7 began in a seated position and the time taken to stand up five times was recorded. Participants  
18  
19  
20 8 were asked to keep their arms folded across their chest throughout the test.  
21  
22 9

23  
24 10 Gait assessment took place using a 4.88 m computerised walkway with embedded pressure  
25  
26  
27 11 sensors (GAITRite®, CIR Systems Inc, New York, USA). Participants performed two walks at their  
28  
29 12 normal pace followed by two walks under cognitive dual task conditions and manual dual task  
30  
31 13 conditions. The cognitive task was to recite alternate letters of the alphabet (A-C-E, etc). The  
32  
33  
34 14 manual task was to carry a glass of water filled to 7 mm from the top. Participants started and  
35  
36  
37 15 finished 2.5 m before and after the walkway to allow for acceleration and deceleration. The two  
38  
39 16 walks in each condition were combined to give mean UGS, mean cognitive dual task gait speed  
40  
41 17 (CGS) and mean manual dual task gait speed (MGS).  
42  
43  
44 18

### 45 46 19 **Statistical analysis**

47  
48  
49 20 This analysis includes participants who completed and had valid scores for baseline and repeat  
50  
51 21 assessments for each of the mobility tests (Figure 1). Missing data was not imputed. To look for  
52  
53  
54 22 practice effects, rater effects and time of day effects, mean mobility performance scores were  
55  
56 23 compared (i) between baseline and repeat assessments, (ii) between raters, and (iii) at different  
57  
58  
59 24 times of day using paired t-tests.  
60

1  
2  
3 1  
4  
5 2 To estimate reliability, mixed effects regression models were then used to find the variation  
6  
7 3 between and within participants. Baseline/repeat assessment, rater and time of day were included  
8  
9  
10 4 as fixed effects. The standard deviations of the within-person and between-person variance  
11  
12 5 components arising from these models were used to estimate the residual ICC for all measures  
13  
14 6 within this population. The ICC used here is the proportion of total variance not accounted for by  
15  
16  
17 7 within person variation, that is,  $CC = \frac{SD_{Between}^2}{SD_{Between}^2 + SD_{Within}^2}$ . Koo & Li<sup>32</sup> recommend that the 95%  
18  
19  
20 8 confidence interval of the ICC estimate is used to evaluate reliability and also suggest the following  
21  
22  
23 9 guidelines: <0.5 indicates poor reliability, 0.5-0.75 indicates moderate reliability, 0.75-0.90  
24  
25 10 indicates good reliability, >0.90 indicates excellent reliability.

26  
27  
28 11  
29  
30 12 SEM is equivalent to  $SD_{Within}$  the standard deviation of the variance of the test within individuals,  
31  
32  
33 13 assuming no genuine change in function, and so is an absolute measure of test reliability. MDC is  
34  
35 14 the magnitude of observable change required to exceed the anticipated measurement error and  
36  
37  
38 15 within-subject variability. It is calculated by  $\sqrt{2} \times Z \times SD_{Within}$  where  $Z=1.96$  for the 95% limit  
39  
40 16 (that is, 95% of observed differences between pairs of observations will be within this limit given  
41  
42  
43 17 there is no true difference) and  $Z=1.65$  for the 90% limit.

44  
45 18  
46  
47 19 The variability of TUG time and RCS time are related to their magnitude; that is, an individual with  
48  
49  
50 20 a TUG time of 4 s is likely to have a lower absolute variation than someone with a TUG time of 12  
51  
52 21 s. For this reason, we estimate the reliability of TUG and RCS on a log-scale, as errors are more  
53  
54  
55 22 likely to be multiplicative than additive, and TUG is often analysed on a logarithmic scale in  
56  
57 23 epidemiological settings.

1  
2  
3 1 Finally, to test whether our estimate of variation is affected by the length of time between  
4  
5 2 assessments we plotted the absolute difference between baseline and repeat measures against  
6  
7 3 the time between assessments, along with a linear model estimated for this relationship.  
8  
9  
10 4

### 11 5 ***Participant and public involvement***

12  
13  
14 6 This research was done without participant involvement. Participants were not invited to  
15  
16  
17 7 comment on the study design and were not consulted to develop participant relevant outcomes or  
18  
19 8 interpret the results. Participants were not invited to contribute to the writing or editing of this  
20  
21  
22 9 document for readability or accuracy.  
23  
24  
25 10

### 26 11 **Results**

27  
28  
29 12 The median age of the sample was 66 years (range 51-89 years, IQR 61-71 years) and 55.5% were  
30  
31  
32 13 female. The majority of the sample (n=103, 81.8%) rated their own health as excellent, very good  
33  
34 14 or good, 57.8% reported having no history of cardiovascular or chronic conditions while 16.0% had  
35  
36  
37 15 3 or more conditions. Median delay between assessments was 88 days (range 28-141 days, IQR  
38  
39 16 70-104 days). Sixty-one participants had a different nurse at the repeat assessment while 60  
40  
41  
42 17 participants had their assessment at a different time of day.  
43  
44  
45 18

46  
47 19 Table 1 shows the mobility performance scores at baseline and repeat assessments, with different  
48  
49 20 raters and at different times of day, while Table 2 shows the variance components and reliability  
50  
51  
52 21 estimates. In general, this sample was relatively robust with good levels of mobility as evidenced  
53  
54  
55 22 when comparing mean TUG and gait speed performance to normative data for community-  
56  
57 23 dwelling adults in Ireland <sup>1</sup>. Norms for RCS are not available for the Irish population, but average  
58  
59 24 performance was slightly slower than age-matched norms presented elsewhere in the literature <sup>33</sup>  
60

1  
2 1 although wide variation in testing protocols has been recognised<sup>34</sup>. Figure 2 shows the baseline  
3  
4 2 versus repeat scores for each measure, while Figure 3 shows the relationship between the  
5  
6 3 absolute differences between scores and the number of days between assessments. In general,  
7  
8 4 there is little evidence that lag between assessments affects the differences, although for TUG, the  
9  
10 5 difference appears slightly lower with increasing time while for RCS the difference appears slightly  
11  
12 6 greater.  
13  
14  
15  
16  
17  
18

### 19 8 *Timed Up-and-go*

20  
21  
22 9 TUG did not vary between baseline and repeat assessments or by time of day, however there was  
23  
24 10 a significant rater effect with a difference of 1.22 s ( $P<.001$ ) between the two nurses. The  
25  
26 11 between-person SD was 1.31 s. The SEM was 0.75 s, leading to moderate-good reliability in this  
27  
28 12 population (ICC=0.75) and MDC estimates of 1.75 s at the 90% level and 2.08 s at the 95% level.  
29  
30 13 This means that a difference of 1.75-2.08 s between two assessments in the same individual can  
31  
32 14 be expected by chance depending on the confidence interval used and when controlling for all  
33  
34 15 other factors (rater, time between assessments and time of day). Analysis of TUG on a logarithmic  
35  
36 16 scale suggests similar reliability (ICC=0.71), and a SEM of 0.09. The  $MDC_{95}$  of 0.24 for  $\log(TUG)$   
37  
38 17 suggests that a relative change in TUG of up to 27% (the inverse logarithm of 0.24 is 1.27) might be  
39  
40 18 expected by chance in 95% of paired samples. This finding is applicable across the spectrum of  
41  
42 19 baseline TUG scores.  
43  
44  
45  
46  
47  
48  
49  
50

### 51 21 *Repeated chair stands*

52  
53 22 RCS was completed slightly more quickly at the repeat measurement (difference=0.47 s,  $P=.04$ )  
54  
55 23 and when the assessment was carried out by Nurse 1 (difference=1.09 s,  $P<.001$ ) but did not vary  
56  
57 24 with time of day. The ICC was 0.66 and SEM was 1.63 s while MDC was estimated to be 3.80 s at  
58  
59  
60

1  
2  
3 1 the 90% level and 4.52 s at the 95% level. Time to complete RCS was also analysed on the log  
4  
5 2 scale, where reliability was similar (ICC=0.68), SEM was 0.13 and MDC was 0.35 at the 95%  
6  
7 3 confidence level (see Table 2).  
8  
9

#### 10 4 11 12 5 *Usual gait speed*

13  
14 6 UGS did not vary between baseline and repeat assessment or by time of day, however there was a  
15  
16  
17 7 significant rater effect with a difference of 7.36 cm/s ( $P<.001$ ). Reliability was good (ICC=0.88) as  
18  
19 8 the between-person SD (18.65 cm/s) was much higher than the SEM (7.03 cm/s), resulting in a  
20  
21  
22 9 MDC<sub>90</sub> of 16.40 cm/s and MDC<sub>95</sub> of 19.49 cm/s (see Table 2 and Figure 2).  
23  
24

#### 25 10 26 27 11 *Manual dual task gait speed*

28  
29 12 Gait speed became less reliable as the complexity of the dual task conditions increased. MGS was  
30  
31  
32 13 consistent across repeat assessments but varied by rater (difference=4.88 cm/s,  $P=.02$ ) and time of  
33  
34 14 day (difference=3.62 s,  $P=0.03$ ). ICC was lower than was observed for UGS (ICC=0.83), SEM was  
35  
36 15 higher (8.97 cm/s) and consequently so was MDC<sub>90</sub> (20.93 cm/s) and MDC<sub>95</sub> (24.87 cm/s) (see  
37  
38  
39 16 Table 2).  
40  
41

#### 42 17 43 44 18 *Cognitive dual task gait speed*

45  
46 19 CGS did not vary by repeat assessment, rater or time of day, however reliability estimates were  
47  
48  
49 20 poorest out of all gait speed measures (ICC=0.77; SEM=12.53 cm/s; MDC<sub>95</sub>=34.73 cm/s) (see Table  
50  
51 21 2).  
52  
53

54 22  
55  
56  
57  
58  
59  
60

1  
2  
3 1 For all observed rater effects, including those where performance was automatically measured  
4  
5 2 (i.e. with GAITRite), participants completed the mobility tasks more quickly when assessed by  
6  
7 3 Nurse 1.  
8  
9  
10 4

## 11 5 **Discussion**

12 6 We report test-retest reliability, SEM and MDC of commonly used mobility tests in a sample of  
13  
14 7 relatively healthy, community-dwelling Irish adults aged 50 years and older. We found good test-  
15  
16 8 retest reliability for walking speed and motor dual task walking speed and moderate-good  
17  
18 9 reliability for TUG and cognitive dual task walking speed however, the lowest ICC was observed for  
19  
20 10 RCS. These findings contrast to previous studies which reported moderate to excellent reliability  
21  
22 11 for all of these measures<sup>9-11 18 19 21-25 35</sup>. As ICC depends on the distribution of scores within the  
23  
24 12 sample it is estimated in and reflects relative reliability, it is specific to that particular setting and  
25  
26 13 population<sup>8</sup>. Lower reliability here is likely to reflect more homogeneous population  
27  
28 14 representative samples (hence lower between-person standard deviations) compared to clinical  
29  
30 15 samples with varying degrees of impairment.  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40

41 16  
42 17 SEM and MDC provide an indication of absolute reliability. MDC allows the assessor to interpret if  
43  
44 18 an observed change score is above that expected due to measurement error and therefore if it  
45  
46 19 represents a genuine change in performance. In this study, MDC for TUG (2.08 s at the 95% level)  
47  
48 20 is lower than that presented in previous studies of healthy (MDC<sub>95</sub>=4.71 s)<sup>16</sup> and cognitively  
49  
50 21 impaired (MDC<sub>95</sub>=5.88-6.87 s) older adults<sup>14 15</sup> and Parkinson's disease patients (MDC<sub>95</sub>=11 s)<sup>17</sup>.  
51  
52  
53 22 However, reporting variability in TUG as a percentage change in performance rather than in  
54  
55 23 absolute terms may be more appropriate. In contrast, MDC<sub>95</sub> for UGS, MGS and CGS  
56  
57 24 (MDC<sub>95</sub>=19.49-34.76 cm/s) are generally higher than the values estimated in community-dwelling  
58  
59  
60

1  
2 1 healthy adults ( $MDC_{95}=13.6$  cm/s)<sup>22</sup>, community-dwelling and hospitalised fallers ( $MDC_{95}=12.4$ -  
3  
4 15.5 cm/s)<sup>35</sup> and in those post-stroke ( $MDC_{95}=20$  cm/s)<sup>36</sup>. These differences may be due to the  
5  
6  
7 3 position on the performance scale as participants in these studies demonstrated poorer mobility  
8  
9  
10 4 than participants in the SHARE-TILDA study<sup>22 35 36</sup>.

11  
12 5  
13  
14 6 Many longitudinal or intervention based studies vary widely in sample characteristics, co-  
15  
16 morbidity and time intervals between assessments. This makes cross-study comparisons difficult  
17 7  
18 and therefore reliability measures are best estimated for each sample and for groups with specific  
19 8  
20 diagnoses. This study provides guidance on MDC across the range of function in a generally  
21 9  
22 healthy, population-based sample, when measurements are compared weeks or months apart.  
23 10  
24 These estimates should be used when assessing individual changes in mobility performance over  
25 11  
26 this time-scale e.g. when examining the effects of an intervention or patient progression, when  
27 12  
28 calculating required sample sizes for studies using these outcomes or when applying methods to  
29 13  
30 adjust for measurement error in epidemiological studies. Participants in this study were relatively  
31 14  
32 healthy and while acute changes in health and performance can occur even with shorter follow-  
33 15  
34 up, they are unlikely to demonstrate a consistent, genuine change in performance in the time  
35 16  
36 period examined. While using a shorter time period and/or same-day repeated measurements  
37 17  
38 would likely provide higher estimates of reliability, this approach was taken to reflect the variation  
39 18  
40 that is likely to be observed in real-world clinical and research settings over a longer time period.  
41 19  
42  
43  
44  
45  
46  
47  
48  
49  
50

51 21 These results show the significant effect of inter-rater variation even with two highly trained and  
52  
53 experienced research nurses. This suggests that changing rater introduces additional variance in  
54 22  
55 the measures beyond within-participant variation. The effect was observed in the GAITRite®  
56 23  
57 assessment as well as stopwatch based tests suggesting that rater differences in reaction time do  
58 24  
59  
60



1  
2  
3 1 not explain this. Both nurses were highly experienced and followed standardised protocols,  
4  
5 2 however one explanation could be that they have different styles of interaction with respondents,  
6  
7 3 which may have impacted on the respondent's understanding of the task, or their motivation and  
8  
9  
10 4 subsequent desire to perform well. This emphasises the importance of providing appropriate  
11  
12 5 training for all raters to ensure that measurements are as accurate and consistent as possible. In  
13  
14 6 an effort to detect and address these differences, studies could examine within-day rater  
15  
16  
17 7 differences on a small number of participants although only a limited number of tests would be  
18  
19 8 feasible to avoid fatigue effects. Where possible, analyses should also be adjusted to account for  
20  
21  
22 9 differences between the raters conducting the assessments.  
23  
24  
25 10

### 26 27 11 ***Study Strengths and Limitations***

28  
29 12 A strength of this study is the population-based sample of relatively healthy middle-aged and older  
30  
31 13 adults used in the analysis. In addition, our estimates of reliability remove time of day and rater  
32  
33 14 effects. For measures that are skewed, a different MDC may be required depending on whether  
34  
35 15 performance is at the higher or lower ends of the spectrum. To account for this, we represent  
36  
37 16 relevant findings on the multiplicative (logarithmic) scale and the additive scale. Although a  
38  
39 17 stopwatch is the easiest and most cost effective way to measure gait speed, the GAITRite® mat is  
40  
41 18 frequently used in research. Therefore, this analysis provides useful guidance on data obtained  
42  
43 19 using simple and more complex instruments. However, there are also a number of limitations in  
44  
45 20 this study. Participants were not asked to restrict their exercise levels, activities or medications  
46  
47 21 before the assessments, all of which could contribute to measurement variation. While the  
48  
49 22 participants did not report any injuries that prevented them from doing the tests, it is also possible  
50  
51 23 that they may have had a low level injury or have been recovering from an injury at either  
52  
53 24 assessment which may account for some of the within-subject variation observed. It is possible  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 1 that underlying mobility among our participants genuinely varied between assessments rather  
4  
5 2 than observed differences representing measurement error or transient factors. However, if this  
6  
7 3 was the case for a significant number of participants, then we would expect to see the differences  
8  
9  
10 4 increase with increasing number of days between assessments. In fact, there was little evidence  
11  
12 5 that the time between assessments contributed to the differences observed.  
13  
14  
15 6

## 17 7 **Conclusion**

18  
19 8 Gait speed obtained during normal walking conditions and when completing a manual dual task  
20  
21  
22 9 are repeatable when performed at time intervals of several weeks to months, with lower reliability  
23  
24 10 observed for the cognitive dual walk, TUG and RCS. There is also a potentially large effect of rater,  
25  
26  
27 11 even for measures that are automatically measured. The estimates of MDC are presented for a  
28  
29 12 population based sample of relatively healthy middle-aged and older Irish adults and can be used  
30  
31  
32 13 to assess changes in performance in individuals drawn from comparable populations. Similar  
33  
34 14 robust reliability studies are recommended to inform the use and interpretation of repeated  
35  
36  
37 15 assessments in other populations such as those with specific co-morbidities. Additional analysis  
38  
39 16 using anchor-based approaches could be used to examine if these changes are of clinical  
40  
41  
42 17 importance.  
43  
44 18

## 46 19 **Author contributions:**

47  
48  
49 20 Substantial contributions to the conception or design of the work; or the acquisition, analysis, or  
50  
51 21 interpretation of data for the work – OD, GS, AB-S, RAK; Drafting the work or revising it critically  
52  
53  
54 22 for important intellectual content – OD, GS, AB-S, RAK; Final approval of the version to be  
55  
56 23 published – OD, GS, AB-S, RAK; Agreement to be accountable for all aspects of the work in  
57  
58  
59  
60

1  
2  
3 1 ensuring that questions related to the accuracy or integrity of any part of the work are  
4  
5 2 appropriately investigated and resolved – OD, GS, AB-S, RAK.  
6  
7 3  
8  
9 4

## 11 5 **References**

- 12  
13  
14 6 1. Kenny RA, Coen RF, Frewen J, et al. Normative values of cognitive and physical function in older  
15  
16  
17 7 adults: findings from the Irish Longitudinal Study on Ageing. *Journal of the American*  
18  
19 8 *Geriatric Society* 2013;61(2):S279-S90.  
20  
21  
22 9 2. Abellan Van Kan G, Rolland Y, Andrieu S, et al. Gait speed at usual pace as a predictor of adverse  
23  
24 10 outcomes in community-dwelling older people an International Academy on Nutrition and  
25  
26 11 Aging (IANA) Task Force. *Journal of Nutrition Health and Aging* 2009;13(10):881-89. doi:  
27  
28 12 10.1007/s12603-009-0246-z  
29  
30  
31 13 3. Cooper R, Kuh D, Hardy R, et al. Objectively measured physical capability levels and mortality:  
32  
33 14 systematic review and meta-analysis. *BMJ* 2010;341(341):c4467. doi: 10.1136/bmj.c4467  
34  
35  
36 15 4. Cooper R, Kuh D, Cooper C, et al. Objective measures of physical capability and subsequent  
37  
38 16 health: a systematic review. *Age Ageing* 2011;40(1):14-23.  
39  
40  
41 17 5. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between  
42  
43 18 reproducibility and responsiveness. *Qual Life Res* 2001;10(7):571-8.  
44  
45  
46 19 6. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30(1):1-  
47  
48 20 15.  
49  
50  
51 21 7. Glymour MM, Weuve J, Berkman LF, et al. When is baseline adjustment useful in analyses of  
52  
53 22 change? An example with education and cognitive change. *Am J Epidemiol*  
54  
55 23 2005;162(3):267-78.  
56  
57  
58  
59  
60

- 1  
2  
3 1 8. Rydwik E, Bergland A, Forsen L, et al. Investigation into the reliability and validity of the  
4  
5 2 measurement of elderly people's clinical walking speed: a systematic review. *Physiother*  
6  
7 3 *Theory Pract* 2012;28(3):238-56.
- 8  
9  
10 4 9. Griswold D, Rockwell K, Killa C, et al. Establishing the reliability and concurrent validity of  
11  
12 5 physical performance tests using virtual reality equipment for community-dwelling healthy  
13  
14 6 elders. *Disability and Rehabilitation* 2014;25:1-5. doi: doi:10.3109/09638288.2014.952451
- 15  
16  
17 7 10. Regterschot GRH, Zhang W, Baldus H, et al. Test–retest reliability of sensor-based sit-to-stand  
18  
19 8 measures in young and older adults. *Gait Posture* 2014;40(1):220-24. doi:  
20  
21  
22 9 <http://dx.doi.org/10.1016/j.gaitpost.2014.03.193>
- 23  
24  
25 10 11. Ng SS, Hui-Chan CW. The Timed Up & Go Test: Its Reliability and Association With Lower-Limb  
26  
27 11 Impairments and Locomotor Capacities in People With Chronic Stroke. *Arch Phys Med*  
28  
29 12 *Rehabil* 2005;86(8):1641-47. doi: 10.1016/j.apmr.2005.01.011
- 30  
31  
32 13 12. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail  
33  
34 14 elderly persons. *J Am Geriatr Soc* 1991;39(2):142-48. [published Online First: PMID:  
35  
36 15 1991946 ]
- 37  
38  
39 16 13. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-  
40  
41 17 Dwelling Older Adults Using the Timed Up & Go Test. *Phys Ther* 2000;80(9):896-903.
- 42  
43  
44 18 14. Blankevoort CG, van Heuvelen MJ, Scherder EJ. Reliability of six physical performance tests in  
45  
46 19 older people with dementia. *Phys Ther* 2013;93(1):69-78.
- 47  
48  
49 20 15. Ries JD, Echternach JL, Nof L, et al. Test-retest reliability and minimal detectable change scores  
50  
51 21 for the timed "up & go" test, the six-minute walk test, and gait speed in people with  
52  
53  
54 22 Alzheimer disease. *Phys Ther* 2009;89(6):569-79.
- 55  
56  
57  
58  
59  
60

- 1  
2  
3 1 16. Mangione KK, Craik RL, McCormick AA, et al. Detectable Changes in Physical Performance  
4  
5 2 Measures in Elderly African Americans. *Phys Ther* 2010;90(6):921-27. doi:  
6  
7 3 10.2522/ptj.20090363  
8  
9  
10 4 17. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and  
11  
12 5 ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease  
13  
14 6 rating scale in people with parkinsonism. *Phys Ther* 2008;88(6):733-46.  
15  
16  
17 7 18. Goldberg A, Chavis M, Watkins J, et al. The five-times-sit-to-stand test: validity, reliability and  
18  
19 8 detectable change in older females. *Aging Clin Exp Res* 2012;24(4):339-44.  
20  
21  
22 9 19. Hollman JH, Childs KB, McNeil ML, et al. Number of strides required for reliable measurements  
23  
24 10 of pace, rhythm and variability parameters of gait during normal and dual task walking in  
25  
26 11 older individuals. *Gait Posture* 2010;32(1):23-28. doi: 10.1016/j.gaitpost.2010.02.017  
27  
28  
29 12 20. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
30  
31 13 in community-dwelling and hospitalized older fallers. *Gait Posture* (38(4):1010-4)  
32  
33  
34 14 21. Brach JS, Perera S, Studenski S, et al. The Reliability and Validity of Measures of Gait Variability  
35  
36 15 in Community-Dwelling Older Adults. *Arch Phys Med Rehabil* 2008;89(12):2293-96. doi:  
37  
38 16 10.1016/j.apmr.2008.06.010  
39  
40  
41 17 22. Goldberg A, Schepens S. Measurement error and minimum detectable change in 4-meter gait  
42  
43 18 speed in older adults. *Aging Clin Exp Res* 2011;23(5-6):406-12.  
44  
45  
46 19 23. Menz HB, Lord SR, St George R, et al. Walking stability and sensorimotor function in older  
47  
48 20 people with diabetic peripheral neuropathy. *Arch Phys Med Rehabil* 2004;85(2):245-52.  
49  
50 21 doi: 10.1016/j.apmr.2003.06.015  
51  
52  
53 22 24. Van Iersel MB, Benraad CEM, Olde Rikkert MGM. Validity and reliability of quantitative gait  
54  
55 23 analysis in geriatric patients with and without dementia. *J Am Geriatr Soc*  
56  
57 24 2007;55(4):632-34. doi: 10.1111/j.1532-5415.2007.01130.x  
58  
59  
60

- 1  
2  
3 1 25. Paterson KL, Hill KD, Lythgo ND, et al. The reliability of spatiotemporal gait data for young and  
4  
5 2 older women during continuous overground walking. *Arch Phys Med Rehabil*  
6  
7 3 2008;89(12):2360-5.  
8  
9  
10 4 26. Börsch-Supan A, Brandt M, Hunkler C, et al. Data Resource Profile: The Survey of Health,  
11  
12 5 Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*  
13  
14 6 2013;42(4):992-1001.  
15  
16  
17 7 27. UCD Geary Institute and Irish Centre for Social Gerontology N. SHARE Ireland - Survey of  
18  
19 8 Health, Ageing and Retirement in Europe [Internet]. Available from:  
20  
21 [http://gearyucdie/share/fileadmin/user\\_upload/shareresults/Share\\_Wave1\\_Resultspdf](http://gearyucdie/share/fileadmin/user_upload/shareresults/Share_Wave1_Resultspdf)  
22 9  
23 2008  
24 10  
25  
26  
27 11 28. Feeney J, Savva GM, O'Regan C, et al. Measurement Error, Reliability, and Minimum  
28  
29 12 Detectable Change in the Mini-Mental State Examination, Montreal Cognitive Assessment,  
30  
31 13 and Color Trails Test among Community Living Middle-Aged and Older Adults. *J Alzheimers*  
32  
33 *Dis* 2016;53(3):1107-14. doi: 10.3233/jad-160248 [published Online First: 2016/06/04]  
34 14  
35  
36  
37 15 29. Finucane C, Savva GM, Kenny RA. Reliability of orthostatic beat-to-beat blood pressure tests:  
38  
39 16 implications for population and clinical studies. *Clin Auton Res* 2017;27(1):31-39. doi:  
40  
41 17 10.1007/s10286-016-0393-3 [published Online First: 2017/01/14]  
42  
43  
44 18 30. Cronin H, O'Regan C, Finucane C, et al. Health and aging: development of the Irish Longitudinal  
45  
46 19 Study on Ageing health assessment. *J Am Geriatr Soc* 2013;61(2):12197.  
47  
48  
49 20 31. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing  
50  
51 21 lower extremity function: association with self-reported disability and prediction of  
52  
53 22 mortality and nursing home admission. *J Gerontol* 1994;49(2):M85-94.  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 1 32. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for  
4  
5 2 Reliability Research. *J Chiropr Med* 2016;15(2):155-63. doi: 10.1016/j.jcm.2016.02.012  
6  
7 3 [published Online First: 2016/03/31]  
8  
9  
10 4 33. Bohannon RW. Reference values for the five-repetition sit-to-stand test: a descriptive meta-  
11  
12 5 analysis of data from elders. *Perceptual and motor skills* 2006;103(1):215-22. doi:  
13  
14 6 10.2466/pms.103.1.215-222 [published Online First: 2006/10/14]  
15  
16  
17 7 34. Mehmet H, Yang AWH, Robinson SR. What is the optimal chair stand test protocol for older  
18  
19 8 adults? A systematic review. *Disability and Rehabilitation* 2019:1-8. doi:  
20  
21 9 10.1080/09638288.2019.1575922  
22  
23  
24 10 35. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables  
25  
26 11 in community-dwelling and hospitalized older fallers. *Gait Posture* 2013;38(4):1010-4.  
27  
28  
29 12 36. Lewek MD, Randall EP. Reliability of spatiotemporal asymmetry during overground walking for  
30  
31 13 individuals following chronic stroke. *J Neurol Phys Ther* 2011;35(3):116-21.  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 1. Mobility performance scores obtained at baseline and repeat assessments, with different raters and at different times of day.

	Assessment		Rater <sup>a</sup>		Time of day <sup>b</sup>	
	Baseline	Repeat	Nurse 1	Nurse 2	Test AM	Test PM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
TUG (s)	8.88 (1.39)	8.87 (1.54)	8.13 (1.20)	9.35 (1.51)***	8.83 (1.49)	8.69 (1.25)
log(TUG)	2.17 (0.02)	2.17 (0.01)	2.08 (0.02)	2.22 (0.02)***	2.16 (0.02)	2.15 (0.02)
RCS (s)	12.49 (2.87)	12.02 (2.48)*	11.80 (2.27)	12.89 (2.88)***	12.17 (2.99)	12.00 (2.46)
logRCS	2.50 (0.22)	2.46 (0.21)*	2.45 (0.20)	2.53 (0.24)**	2.47 (0.24)	2.46 (0.22)
UGS (cm/s)	137.95 (20.21)	138.20 (19.32)	145.82 (18.94)	138.46 (17.85)***	137.62 (17.68)	137.74 (17.38)
MGS (cm/s)	116.76 (21.84)	118.71 (19.93)	123.07 (18.95)	118.07 (20.45)**	117.86 (19.85)	122.19 (17.21)*
CGS (cm/s)	115.23 (24.08)	115.15 (25.21)	118.29 (25.24)	117.40 (20.99)	117.45 (24.01)	118.84 (20.18)

Notes: TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task gait speed

<sup>a</sup> Rater scores are calculated only among participants who changed rater at the repeat assessment

<sup>b</sup> Time of day scores are calculated only among participants who changed time of day at the repeat assessment.

\*  $P < 0.05$ ; \*\*  $P < 0.01$ ; \*\*\*  $P < 0.001$



Table 2. Variance and reliability estimates for all mobility tests.

	<b>SD<sub>between</sub> (95% CI)</b>	<b>SEM (95% CI)</b>	<b>ICC (95% CI)</b>	<b>MDC<sub>90</sub></b>	<b>MDC<sub>95</sub></b>
TUG (s)	1.31 (1.12-1.52)	0.75 (0.66-0.85)	0.75 (0.66-0.82)	1.75	2.08
logTUG	0.13 (0.11-0.15)	0.09 (0.08-0.10)	0.71 (0.61-0.79)	0.2	0.24
RCS (s)	2.29 (1.93-2.70)	1.63 (1.43-1.86)	0.66 (0.55-0.76)	3.8	4.52
logRCS	0.18 (0.16-0.22)	0.13 (0.11-0.14)	0.68 (0.57-0.77)	0.29	0.35
UGS (cm/s)	18.65 (16.34-21.29)	7.03 (6.20-7.98)	0.88 (0.83-0.91)	16.4	19.49
MGS (cm/s)	19.57 (17.04-22.46)	8.97 (7.90-10.19)	0.83 (0.76-0.88)	20.93	24.87
CGS (cm/s)	22.73 (19.62-26.34)	12.53 (10.99-14.28)	0.77 (0.68-0.83)	29.24	34.73

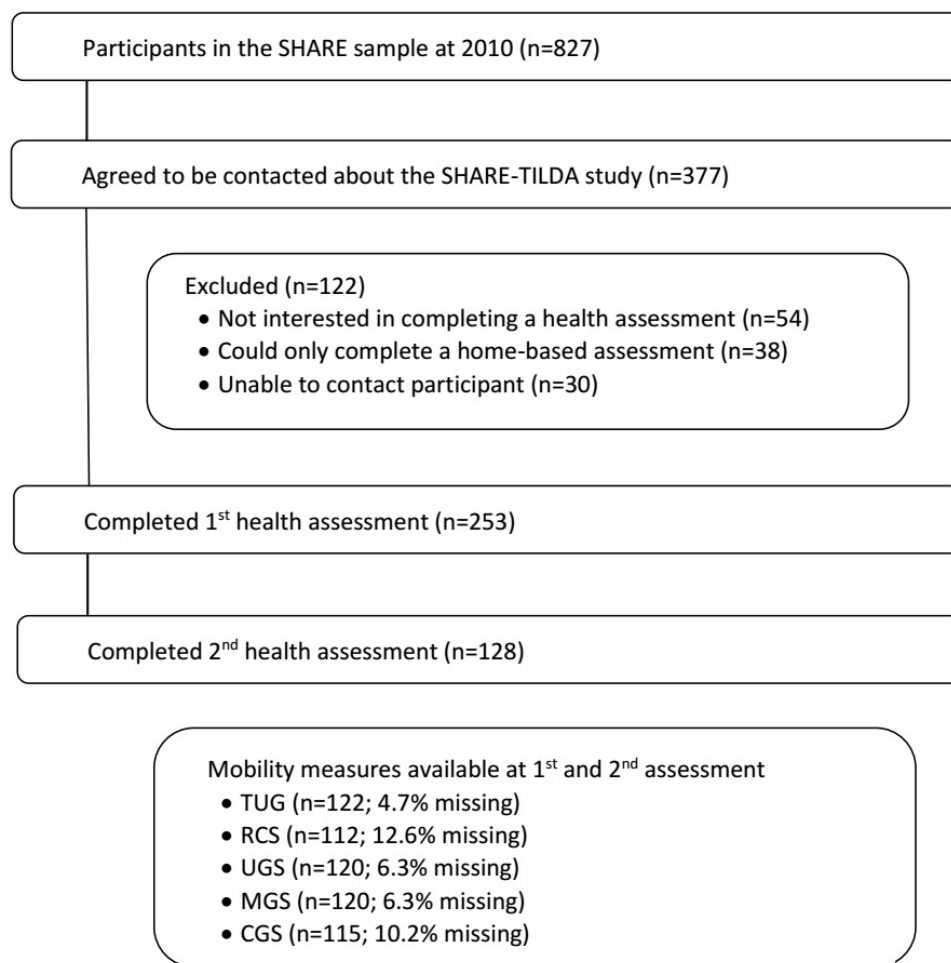
Notes: SEM, standard error of the measurement; TUG, Timed Up-and-Go; RCS, repeated chair stands; UGS, usual gait speed; MGS, manual dual task gait speed; CGS, cognitive dual task speed; ICC, intra-class correlation; MDC, minimum detectable change

1  
2 Figure 1: Exclusion criteria used to establish eligible participants for this analysis.  
3

4 Note: CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair  
5 stands; SHARE, Survey for Health Ageing and Retirement in Europe; TILDA, The Irish Longitudinal  
6 Study on Ageing; TUG, Timed Up-and-Go; UGS, usual gait speed.  
7  
8  
9  
10  
11  
12  
13  
14

15 Figure 2. Scatter plots showing the relationship between baseline (measure 1) and repeat  
16 (measure 2) scores for repeated chair stands (RCS), Timed Up-and-Go (TUG), and gait speed under  
17 normal conditions, with a cognitive dual task and a manual dual task. Solid line represents equality  
18 between the two measures.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

29 Figure 3. The absolute difference between the initial and repeat score for each measure (vertical  
30 axis) plotted against the days between assessments. Lines represent linear regression models with  
31 95% confidence bands.  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Exclusion criteria used to establish eligible participants for this analysis.

245x247mm (96 x 96 DPI)

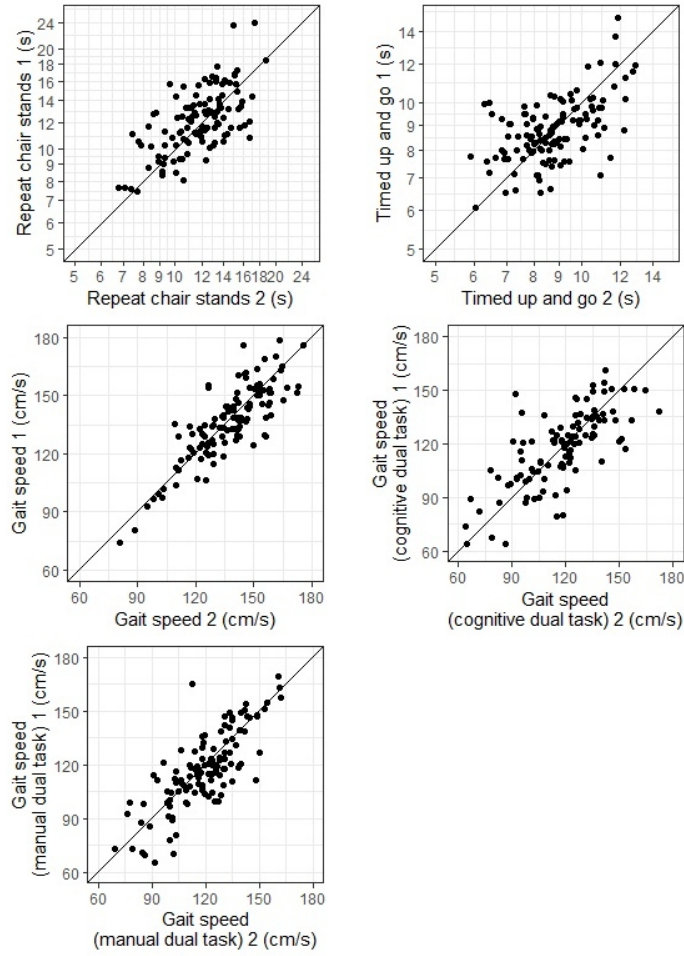


Figure 2

190x254mm (96 x 96 DPI)

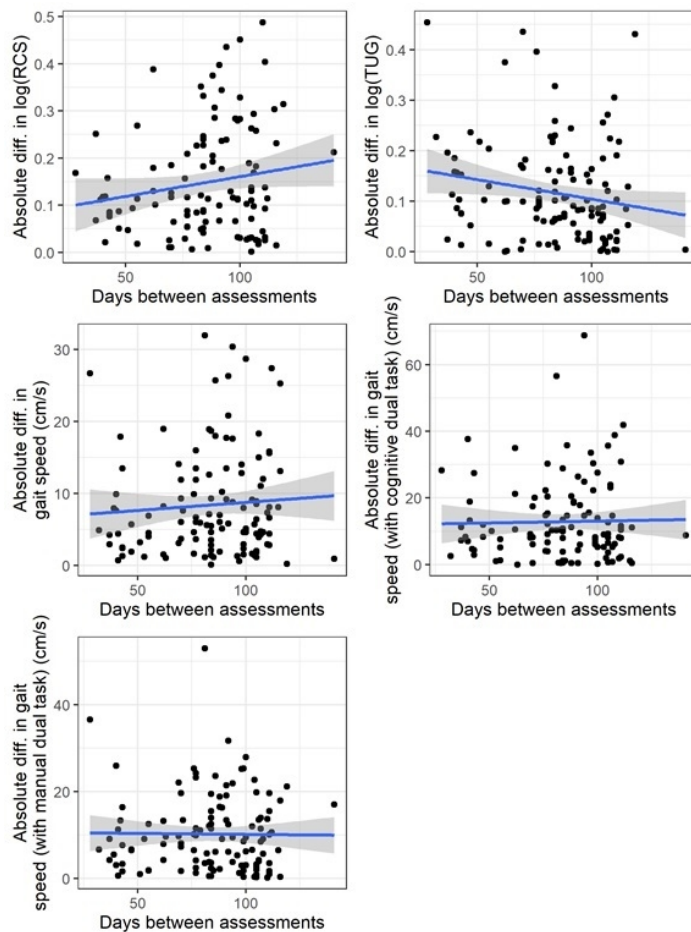


Figure 3

216x203mm (96 x 96 DPI)

1 STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation	Location
<b>Title and abstract</b>	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	P1, P2
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	P2
<b>Introduction</b>			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	P5-7
Objectives	3	State specific objectives, including any prespecified hypotheses	P6
<b>Methods</b>			
Study design	4	Present key elements of study design early in the paper	P7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	P7-8
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	P7-8
		(b) For matched studies, give matching criteria and number of exposed and unexposed	-
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	P8-10
Data sources/measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	P8-10
Bias	9	Describe any efforts to address potential sources of bias	P10
Study size	10	Explain how the study size was arrived at	P7-8
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	P10-11
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	P10-11
		(b) Describe any methods used to examine subgroups and interactions	P10-11
		(c) Explain how missing data were addressed	P10
		(d) If applicable, explain how loss to follow-up was addressed	-
		(e) Describe any sensitivity analyses	-
<b>Results</b>			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	P8, 12, Fig 1
		(b) Give reasons for non-participation at each stage	P8
		(c) Consider use of a flow diagram	Fig 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	P12
		(b) Indicate number of participants with missing data for each variable of interest	Fig 1
		(c) Summarise follow-up time (eg, average and total amount)	P12
Outcome data	15*	Report numbers of outcome events or summary measures over time	P23
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear	P12, 14, 24

		which confounders were adjusted for and why they were included	
		(b) Report category boundaries when continuous variables were categorized	-
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	-
<b>Discussion</b>			
Key results	18	Summarise key results with reference to study objectives	P12-14
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	P17
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	P14-17
Generalisability	21	Discuss the generalisability (external validity) of the study results	P15-17
<b>Other information</b>			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	P4

\*Give information separately for exposed and unexposed groups.

**Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at <http://www.strobe-statement.org>.