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

Orna A Donoghue,1 George M Savva,2 Axel Börsch-Supan,3 Rose Anne Kenny1,4

ABSTRACT

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

Design Test–retest study in a population representative sample.

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

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

Interventions Not applicable.

Outcome measures Participants performed timed up-and-go (TUG), repeated chair stands (RCS) and walking speed tests administered by one of two raters. Repeat assessments were conducted 1–4 months later. Participants were randomised with respect to a change in time (morning, afternoon) and whether the rater was changed between assessments. Within and between-participant variance for each measure was estimated using mixed-effects models. Intraclass correlation (ICC), SE of measurement and minimum detectable change (MDC) were reported.

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

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

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.1 These measures are good predictors of falls, disability, cognitive decline and mortality.2-4 To be useful, they also need to be reliable (consistent when measured on several occasions and when there is no change in an individual’s underlying performance) and responsive (able to detect a change when there is one).5 Good reliability allows changes in measurements to be tracked over time.6 However, all tests are subject to measurement error due to within-subject, intertrial
and inter-rater effects. They are also liable to day-to-day variation due to patient-level factors that do not reflect the underlying risk status that they are attempting to measure. 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 status and is not due to measurement error or a transient effect. In research settings, unreliable measures can lead to regression dilution bias or false positive associations when testing predictors of longitudinal change. To account for this, several measures of relative reliability, that is, intra-class correlation (ICC), and absolute reliability, that is, SE of measurement (SEM) and minimum detectable change (MDC), are often reported.

SEM is the SD 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 to 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 day-to-day fluctuations in these measures, for example, due to acute illness or day-to-day variation, add error to these outcomes along with measurement error associated with the instruments themselves. Hence when comparing measures over longer time periods, that is, years or decades typical of epidemiological research, it is important to know how well single measures of physical and cognitive function reflect the underlying health status of the participant, net of any factors that might cause a short-term fluctuation. Therefore, we tested the concordance between pairs of measures between one and 4 months apart, to estimate the error association with both measurement and day-to-day fluctuation in each measure. Understanding natural variation in outcomes over 1–4 months is also essential when planning clinical trials with follow-up time in this range, since this is the natural variation against which any treatment effect would be compared.

**METHODS**

**Participants**

Participants were a subsample from the Survey of Health, Ageing and Retirement in Europe (SHARE), a longitudinal, cross-national study on health, socioeconomic status and social and family networks of more than 80,000 individuals aged 50 years and over across Europe. The SHARE-Ireland sample (n=1119) was recruited in Ireland between 2006 and 2007 with a response rate of 55%. A collaboration between SHARE-Ireland and The Irish Longitudinal Study on Ageing (TILDA) was established to understand the measurement properties of a comprehensive health assessment among a representative sample of the European population. Reliability of cognitive measures and blood pressure dynamics based on this sample have been published previously.

The extant SHARE-Ireland cohort at 2010 (n=827) was contacted and invited to take part in a health assessment that included the same tests and followed the same protocols as those used by TILDA. The health assessment was delivered to the SHARE-Ireland participants by TILDA research nurses within the TILDA health assessment centre based at Trinity College Dublin. Initial contact was made by post and followed up by telephone between September 2011 and March 2012, with 377 participants consenting to receive further information about the study. Of these, 253 agreed to an initial health assessment (see figure 1).

**Health assessments and interview**

The full health assessment included a 3-hour battery of tests assessing cognitive function, gait and mobility, cardiovascular function and vision. Health assessments

---

Participants in the SHARE sample at 2010 (n=827)
Agreed to be contacted about the SHARE-TILDA study (n=377)

Excluded (n=122)
- Not interested in completing a health assessment (n=54)
- Could only complete a home-based assessment (n=58)
- Unable to contact participant (n=30)

Completed 1st health assessment (n=253)
Completed 2nd health assessment (n=128)

Mobility measures available at 1st and 2nd assessment
- TUG (n=122; 4.7% missing)
- RCS (n=112; 12.6% missing)
- UGS (n=120; 6.3% missing)
- MGS (n=120; 6.3% missing)
- CGS (n=115; 10.2% missing)

Repeat assessments were arranged to distinguish within-person variation from variation caused by changing rater or time of day. The same research nurse conducted the baseline and repeat assessments for half of the participants while another nurse conducted the repeat assessment for the other half of the participants. Time of day when the assessment took place (morning or afternoon) was also changed for half of the participants. Change of rater, change of time of day and delay between assessments (dichotomised at the median) were randomised using a minimisation routine designed to achieve balance across these covariates, as well as the age group and sex of participants. Other factors that could influence performance, for example, health assessment protocols, assessment location, equipment were held constant across both assessments.

**Physical performance tests**

Participants completed several mobility tests—TUG, RCS and gait assessments in single and dual task conditions. TUG, which is a common functional mobility test, was completed once using walking aids if required. The time taken to rise from the chair (seat height 46 cm), walk 3 m at normal pace, turn around, walk back to the chair and sit down again was recorded using a stopwatch. RCS is an indicator of mobility and lower limb muscular endurance. Participants began in a seated position and the time taken to stand up five times was recorded. Participants were asked to keep their arms folded across their chest throughout the test.

Gait assessment took place using a 4.88 m computerised walkway with embedded pressure sensors (GAITRite, CIR Systems, New York, USA). Participants performed two walks at their normal pace, followed by two walks under cognitive dual task conditions and manual dual task conditions. The cognitive task was to recite alternate letters of the alphabet (A–C–E, etc). The manual task was to carry a glass of water filled to 7 mm from the top. Participants started and finished 2.5 m before and after the walkway to allow for acceleration and deceleration. The two walks in each condition were combined to give mean UGS, mean cognitive dual task gait speed (CGS) and mean manual dual task gait speed (MGS).

**Statistical analysis**

This analysis includes participants who completed and had valid scores for baseline and repeat assessments for each of the mobility tests (figure 1). Missing data were not imputed. To look for practice effects, rater effects and time of day effects, mean mobility performance scores were compared (1) between baseline and repeat assessments, (2) between raters and (3) at different times of day using paired t-tests.

To estimate reliability, mixed-effects regression models were then used to find the variation between and within participants. Baseline/repeat assessment, rater and time of day were included as fixed effects. The SD of the within-person and between-person variance components arising...
from these models were used to estimate the residual ICC for all measures within this population. The ICC used here is the proportion of total variance not accounted for by within person variation, that is, $ICC = \frac{SD_{Between}^2}{SD_{Between}^2 + SD_{Within}^2}$.

Koo and Li recommend that the 95% CI of the ICC estimate is used to evaluate reliability and also suggest the following guidelines: <0.5 indicates poor reliability, 0.5–0.75 indicates moderate reliability, 0.75–0.90 indicates good reliability and >0.90 indicates excellent reliability.

SEM is equivalent to $SD_{Within}$, the SD of the variance of the test within individuals, assuming no genuine change in function, and so is an absolute measure of test reliability. MDC is the magnitude of observable change required to exceed the anticipated measurement error and within-subject variability. It is calculated by $\sqrt{2 \times Z^2 \times SD_{Within}}$, where $Z=1.96$ for the 95% limit (ie, 95% of observed differences between pairs of observations will be within this limit given there is no true difference) and $Z=1.65$ for the 90% limit.

The variabilities of TUG time and RCS time are related to their magnitude, that is, an individual with a TUG time of 4s is likely to have a lower absolute variation than someone with a TUG time of 12s. For this reason, we estimate the reliability of TUG and RCS on a log-scale, as errors are more likely to be multiplicative than additive, and TUG is often analysed on a logarithmic scale in epidemiological settings.

Finally, to test whether our estimate of variation is affected by the length of time between assessments we plotted the absolute difference between baseline and repeat measures against the time between assessments, along with a linear model estimated for this relationship.

**Participant and public involvement**

This research was done without participant involvement. Participants were not invited to comment on the study design and were not consulted to develop participant relevant outcomes or interpret the results. Participants were not invited to contribute to the writing or editing of this document for readability or accuracy.

**RESULTS**

The median age of the sample was 66 years (range 51–89 years, IQR 61–71 years) and 55.5% were female. The majority of the sample (n=103, 81.8%) rated their own health as excellent, very good or good, 57.8% reported having no history of cardiovascular or chronic conditions while 16.0% had three or more conditions. Median delay between assessments was 88 days (range 28–141 days, IQR 70–104 days). Sixty-one participants had a different nurse at the repeat assessment while 60 participants had their assessment at a different time of day.

**Table 1** shows the mobility performance scores at baseline and repeat assessments, with different raters and at different times of day, while **table 2** shows the variance components and reliability estimates. In general, this sample was relatively robust with good levels of mobility as evidenced when comparing mean TUG and gait speed performance to normative data for community-dwelling adults in Ireland. Norms for RCS are not available for the Irish population, but average performance was slightly slower than age-matched norms presented elsewhere in the literature although wide variation in testing protocols has been recognised. Figure 2 shows the baseline vs repeat scores for each measure, while figure 3 shows the relationship between the absolute differences between scores and the number of days between assessments. In general, there is little evidence that lag between assessments affects the differences, although for TUG, the difference appears slightly lower with increasing time while for RCS the difference appears slightly greater.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mobility performance scores obtained at baseline and repeat assessments, with different raters and at different times of day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assessment</strong></td>
<td><strong>Rater†</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Baseline Mean (SD)</strong></td>
</tr>
<tr>
<td>TUG (s)</td>
<td>8.88 (1.39)</td>
</tr>
<tr>
<td>Log(TUG)</td>
<td>2.17 (0.02)</td>
</tr>
<tr>
<td>RCS (s)</td>
<td>12.49 (2.87)</td>
</tr>
<tr>
<td>LogRCS</td>
<td>2.50 (0.22)</td>
</tr>
<tr>
<td>UGS (cm/s)</td>
<td>137.95 (20.21)</td>
</tr>
<tr>
<td>MGS (cm/s)</td>
<td>116.76 (21.84)</td>
</tr>
<tr>
<td>CGS (cm/s)</td>
<td>115.23 (24.08)</td>
</tr>
</tbody>
</table>

*P<0.05, **P<0.01, ***P<0.001.
†Rater scores are calculated only among participants who changed rater at the repeat assessment.
‡Time of day scores are calculated only among participants who changed time of day at the repeat assessment.
CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair stands; TUG, timed up-and-go; UGS, usual gait speed.
Table 2  Variance and reliability estimates for all mobility tests

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

CGS, cognitive dual task gait speed; ICC, intraclass correlation; MDC, minimum detectable change; MGS, manual dual task gait speed; RCS, repeated chair stands; SEM, SE of measurement; TUG, timed up-and-go; UGS, usual gait speed.

**Timed up-and-go**

TUG did not vary between baseline and repeat assessments or by time of day, however, there was a significant rater effect with a difference of 1.22 s (p<0.001) between the two nurses. The between-person SD was 1.31 s. The SEM was 0.75 s, leading to moderate-good reliability in this population (ICC=0.75) and MDC estimates of 1.75 s at the 90% level and 2.08 s at the 95% level. This means that a difference of 1.75–2.08 s between two assessments in the same individual can be expected by chance depending on the CI used and when controlling for all other factors (rater, time between assessments and time of day). Analysis of TUG on a logarithmic scale suggests similar reliability (ICC=0.71), and an SEM of 0.09. The MDC\textsubscript{95} of 0.24 for log(TUG) suggests that a relative change in TUG of up to 27% (the inverse logarithm of 0.24 is 1.27) might...
be expected by chance in 95% of paired samples. This finding is applicable across the spectrum of baseline TUG scores.

**Repeated chair stands**
RCS was completed slightly more quickly at the repeat measurement (difference=0.47 s, p=0.04) and when the assessment was carried out by nurse 1 (difference=1.09 s, p<0.001) but did not vary 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 the 90% level and 4.52 s at the 95% level. Time to complete RCS was also analysed on the log scale, where reliability was similar (ICC=0.68), SEM was 0.13 and MDC was 0.35 at the 95% confidence level (see table 2).

**Usual gait speed**
UGS did not vary between baseline and repeat assessment or by time of day, however, there was a significant rater effect with a difference of 7.36 cm/s (p<0.001). Reliability was good (ICC=0.88) as the between-person SD (18.65 cm/s) was much higher than the SEM (7.03 cm/s), resulting in an MDC90 of 16.40 cm/s and MDC95 of 19.49 cm/s (see table 2 and figure 2).

**Manual dual task gait speed**
Gait speed became less reliable as the complexity of the dual task conditions increased. MGS was consistent across repeat assessments but varied by rater (difference=4.88 cm/s, p=0.02) and time of day (difference=3.62 s, p=0.03). ICC was lower than was observed for UGS (ICC=0.83), SEM was higher (8.97 cm/s) and consequently so was MDC90 (20.93 cm/s) and MDC95 (24.87 cm/s) (see table 2).

**Cognitive dual task gait speed**
CGS did not vary by repeat assessment, rater or time of day, however, reliability estimates were the poorest out of all gait speed measures (ICC=0.77; SEM=12.53 cm/s; MDC90=34.73 cm/s) (see table 2).

For all observed rater effects, including those where performance was automatically measured (ie, with GAITRite), participants completed the mobility tasks more quickly when assessed by nurse 1.

**DISCUSSION**
We report test–retest reliability, SEM and MDC of commonly used mobility tests in a sample of relatively healthy, community-dwelling Irish adults aged 50 years and older. We found good test–retest reliability for walking speed and motor dual task walking speed and moderate-good reliability for TUG and cognitive dual task walking speed, however, the lowest ICC was observed for RCS. These findings contrast to previous studies which reported moderate to excellent reliability for all of these measures.6–11 18–25 As ICC depends on the distribution of scores within the sample it is estimated in and reflects relative reliability, it is specific to that particular setting and population.8 Lower reliability here is likely to reflect more homogeneous population representative samples (hence lower between-person SD) compared with clinical samples with varying degrees of impairment.

SEM and MDC provide an indication of absolute reliability. MDC allows the assessor to interpret if an observed change score is above that expected due to measurement error and therefore if it represents a genuine change in performance. In this study, MDC for TUG (2.08 s at the 95% level) is lower than that presented in previous studies of healthy (MDC95=4.71 s)16 and cognitively impaired (MDC95=5.88–6.87 s) older adults14 15 and Parkinson’s disease patients (MDC95=11 s).17 However, reporting variability in TUG as a percentage change in performance rather than in absolute terms may be more appropriate. In contrast, MDC95 for UGS, MGS and CGS (MDC95=19.49–34.76 cm/s) are generally higher than the values estimated in community-dwelling healthy adults (MDC95=13.6 cm/s)22 community-dwelling and hospitalised fallers (MDC95=12.4–15.5 cm/s)20 and in those poststroke (MDC95=20 cm/s).35 These differences may be due to the position on the performance scale as participants in these studies demonstrated poorer mobility than participants in the SHARE-TILDA study.20 22 35

Many longitudinal or intervention-based studies vary widely in sample characteristics, comorbidity and time intervals between assessments. This makes cross-study comparisons difficult and therefore reliability measures are best estimated for each sample and for groups with specific diagnoses. This study provides guidance on MDC across the range of function in a generally healthy, population-based sample, when measurements are compared weeks or months apart. These estimates should be used when assessing individual changes in mobility performance over this time scale, for example, when examining the effects of an intervention or patient progression, when calculating required sample sizes for studies using these outcomes or when applying methods to adjust for measurement error in epidemiological studies. Participants in this study were relatively healthy and while acute changes in health and performance can occur even with shorter follow-up, they are unlikely to demonstrate a consistent, genuine change in performance in the time period examined. While using a shorter time period and/or same-day repeated measurements would likely provide higher estimates of reliability, this approach was taken to reflect the variation that is likely to be observed in real-world clinical and research settings over a longer time period.

These results show the significant effect of inter-rater variation even with two highly trained and experienced research nurses. This suggests that changing rater introduces additional variance in the measures beyond within-participant variation. The effect was observed in the GAITRite assessment as well as stopwatch-based tests suggesting that rater differences in reaction time do not explain this. Both nurses were highly experienced and followed standardised protocols, however, one explanation could be that they have different styles of interaction.
with respondents, which may have impacted on the respondent’s understanding of the task, or their motivation and subsequent desire to perform well. This emphasises the importance of providing appropriate training for all raters to ensure that measurements are as accurate and consistent as possible. In an effort to detect and address these differences, studies could examine within-day rater differences on a small number of participants although only a limited number of tests would be feasible to avoid fatigue effects. Where possible, analyses should also be adjusted to account for differences between the raters conducting the assessments.

**Study strengths and limitations**

A strength of this study is the population-based sample of relatively healthy middle-aged and older adults used in the analysis. In addition, our estimates of reliability remove time of day and rater effects. For measures that are skewed, a different MDC may be required depending on whether performance is at the higher or lower ends of the spectrum. To account for this, we represent relevant findings on the multiplicative (logarithmic) scale and the additive scale. Although a stopwatch is the easiest and most cost-effective way to measure gait speed, the GAITRite mat is frequently used in research. Therefore, this analysis provides useful guidance on data obtained using simple and more complex instruments. However, there are also a number of limitations in this study. Participants were not asked to restrict their exercise levels, activities or medications before the assessments, all of which could contribute to measurement variation. While the participants did not report any injuries that prevented them from doing the tests, it is also possible that they may have had a low level injury or have been recovering from an injury at either assessment which may account for some of the within-subject variation observed. It is possible that underlying mobility among our participants genuinely varied between assessments rather than observed differences representing measurement error or transient factors. However, if this was the case for a significant number of participants, then we would expect to see the differences increase with increasing number of days between assessments. In fact, there was little evidence that the time between assessments contributed to the differences observed.

**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 task, 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 comorbidities. Additional analysis using anchor-based approaches could be used to examine if these changes are of clinical importance.

**Acknowledgements** The authors would like to acknowledge the contribution of the participants and members of the TILDA and SHARE teams.

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

**Funding** TILDA received financial support from the Irish Government (Department of Health and Children), the Atlantic Philanthropies and Irish Life plc. The SHARE-TILDA project was funded by the National Institute of Aging (Prime Award Number R21AG04387).

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Ethics approval** Ethical approval for this substudy was obtained from the Faculty of Health Sciences Research Ethics Committee at Trinity College Dublin.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** TILDA considers applications for privileged access to the dataset through an onsite “hot desk” facility based in TILDA (visit www.tilda.ie for further information).

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is appropriately investigated and resolved.

**REFERENCES**