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

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

To cite: Donoghue OA, Savva GM, Börsch-Supan A, *et al*. Reliability, measurement error and minimum detectable change in mobility measures: a cohort study of community-dwelling adults aged 50 years and over in Ireland. *BMJ Open* 2019;**9**:e030475. doi:10.1136/bmjopen-2019-030475

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-030475).

Received 15 March 2019
Revised 04 October 2019
Accepted 15 October 2019

[1]The Irish Longitudinal Study on Ageing (TILDA), University of Dublin Trinity College, Dublin, Ireland
[2]Quadram Institute Bioscience, Norwich, UK
[3]Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy, Munchen, Germany
[4]Mercer's Institute for Successful Ageing, St. James' Hospital, Dublin, Ireland

Correspondence to
Dr Orna A Donoghue;
odonogh@tcd.ie

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.

Strengths and limitations of this study

- This study provides information on the effects of repeat assessments, rater and time of day on test–retest reliability of mobility measures obtained over 1–4 months using a population-based sample of relatively healthy middle-aged and older adults aged ≥50 years in Ireland.
- The use of common tests, such as timed up-and-go, repeated chair stands and GAITRite assessments, makes this analysis relevant for other studies looking at change in mobility.
- Mixed-effects models were used to estimate within and between-participant variance for each measure allowing intraclass correlation and SE of measurement and minimum detectable change (MDC) to be presented, net of fixed effects.
- For some measures, MDC was presented on the multiplicative (logarithmic) scale as well as the natural additive scale to account for skewness and to ensure that findings are applicable across all levels of performance.
- Changes in exercise levels, activities, medications and current injury status could have contributed to measurement variation but these were not measured. However, the fact that the measures did not become less reliable with increasing time since assessments suggests that this does not substantially affect the findings.

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.[7] To account for this, several measures of relative reliability, that is, intraclass correlation (ICC), and absolute reliability, that is, SE of measurement (SEM) and minimum detectable change (MDC), are often reported.[8]

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 variation between members of a sample, and so is not usually comparable or applicable across samples with different levels of heterogeneity.

The within-session and 1-week test–retest reliability of TUG in community-dwelling, older adults is well known, and is known to be high (ICC ≥0.96)[9–11] in various populations as is the inter-rater[12 13] and intra-rater reliability.[12] MDC at the 95% confidence level (MDC_{95}) has been reported to vary between 3.33–6.87 s in healthy and cognitively impaired older adults[14–16] and up to 11 s in Parkinson's disease patients.[17] The within-session test–retest reliability of RCS is also very high (ICC=0.93–0.95),[9 18] however, SEM and MDC for community-dwelling adults are not known.

Walking speed can be measured using stopwatches, timing gates or sensored mats. The test-retest reliability of usual gait speed (UGS) measured using a GAITRite walkway has been reported to be between ICC=0.84 and 0.97 for assessments given up to 2 weeks apart.[19–25] Similar values have been reported for 1-hour test–retest reliability of dual task gait speed (ICC=0.85–0.93).[19 20] Fewer studies have reported SEM or MDC in healthy populations with MDC values of 12.4–13.6 cm/s reported for UGS[20 22] and 15.5 cm/s for dual task gait speed.[20] However, reliability of dual task gait speed may also be dependent on the actual dual task, and therefore, is not comparable across studies unless the same test has been used.

Here, we report the test–retest reliability measured by ICC, SEM and MDC in a pragmatic epidemiological setting. We explore how reliability changes when lag between assessments varies between 1 and 4 months, when rater changes or is held constant, and whether or not time of assessment varies, in a large sample of healthy adults aged 50 and older recruited at random from the population.

In epidemiological settings, these measures are commonly used as proxies for the underlying general cognitive and physical health status of participants around the time of the assessment. Short-term 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.[26] The SHARE-Ireland sample (n=1119) was recruited in Ireland between 2006 and 2007 with a response rate of 55%.[27] 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.[28 29]

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.[30] Health assessments

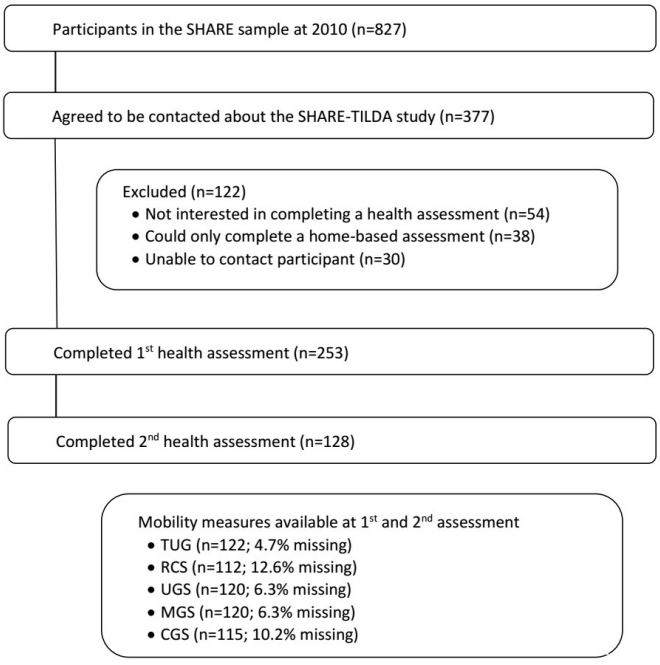

Figure 1 Exclusion criteria used to establish eligible participants for this analysis. CGS, cognitive dual task gait speed; MGS, manual dual task gait speed; RCS, repeated chair stands; SHARE, Survey for Health Ageing and Retirement in Europe; TILDA, The Irish Longitudinal Study on Ageing; TUG, timed up-and-go; UGS, usual gait speed.

were conducted by two highly trained research nurses with approximately 3 years experience delivering these specific tests in the current setting. Training took approximately 1 month and nurses used detailed and standardised health assessment protocols which included clear explanations and demonstrations to ensure consistent instructions were provided to all participants. Nurses also underwent periodic quality control procedures to ensure adherence to the protocols.

A short interview was administered by the nurses before the health assessment to capture information on health, chronic disease, disability, employment, social and financial circumstances. Comorbidity was assessed by asking participants if a doctor had ever told them that they had any of the following conditions: heart attack, high blood pressure, high cholesterol, stroke, diabetes, chronic lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson's disease, cataracts, age-related macular degeneration, Alzheimer's disease and atrial fibrillation. The number of conditions was summed and categorised according to 0, 1, 2 or ≥3 conditions. Participants self-rated their health as excellent, very good, good, fair or poor.

On completing the health assessment, 180 participants were invited to take part in an identical repeat assessment, scheduled after 1–4 months. In total, 128 participants (58 men) agreed to the repeat assessment giving a response rate of 71% (25 refused and 27 were unavailable to attend the repeat assessment within the required time frame).

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,[12] 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.[31] 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, $CC = \frac{SD^2_{Between}}{SD^2_{Between}+SD^2_{Within}}$. Koo and Li[32] 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 $\sqrt{2} \times Z \times SD_{Within}$, 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 4 s is likely to have a lower absolute variation than someone with a TUG time of 12 s. 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.[1] 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[33] although wide variation in testing protocols has been recognised.[34] 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 1 Mobility performance scores obtained at baseline and repeat assessments, with different raters and at different times of day

	Assessment		Rater†		Time of day‡	
	Baseline Mean (SD)	Repeat Mean (SD)	Nurse 1 Mean (SD)	Nurse 2 Mean (SD)	Test AM Mean (SD)	Test PM 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)

*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

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

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_{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

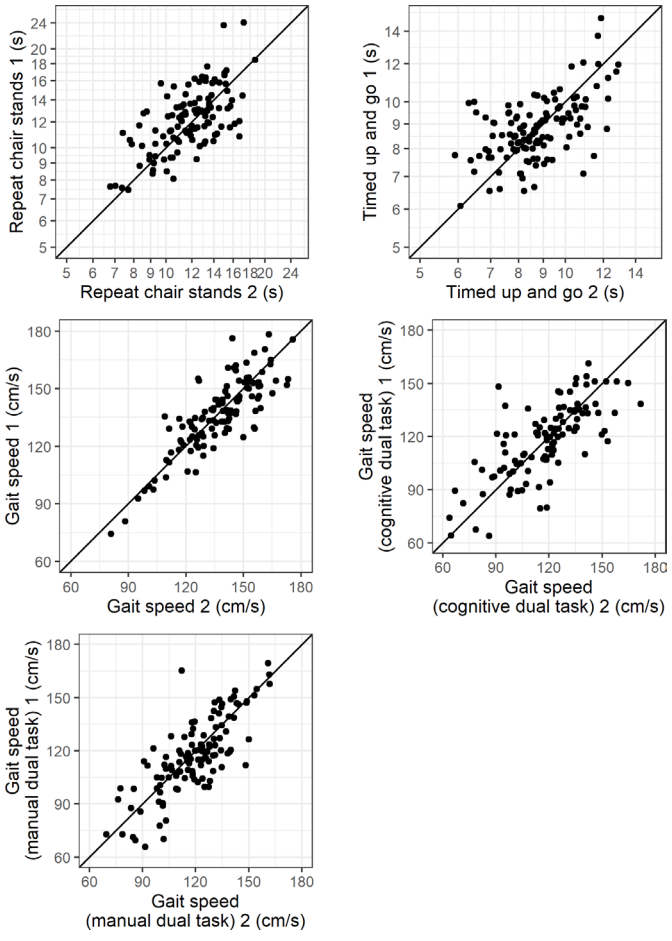

Figure 2 Scatter plots showing the relationship between baseline (measure 1) and repeat (measure 2) scores for repeated chair stands, Timed up-and-go, and gait speed under normal conditions, with a cognitive dual task and a manual dual task. Solid line represents equality between the two measures.

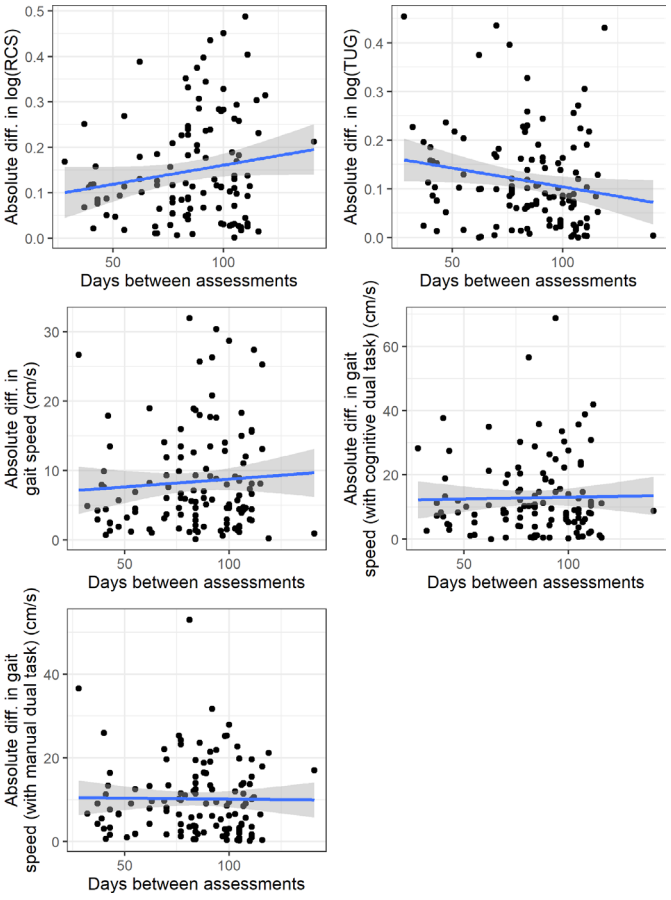

Figure 3 The absolute difference between the initial and repeat score for each measure (vertical axis) plotted against the days between assessments. Lines represent linear regression models with 95% confidence bands. RCS, repeated chair stands; TUG, timed up-and-go.

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 MDC_{90} of 16.40 cm/s and MDC_{95} 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 MDC_{90} (20.93 cm/s) and MDC_{95} (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; MDC_{95}=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.[9–11 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 (MDC_{95}=4.71 s)[16] and cognitively impaired (MDC_{95}=5.88–6.87 s) older adults[14 15] and Parkinson's disease patients (MDC_{95}=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, MDC_{95} for UGS, MGS and CGS (MDC_{95}=19.49–34.76 cm/s) are generally higher than the values estimated in community-dwelling healthy adults (MDC_{95}=13.6 cm/s),[22] community-dwelling and hospitalised fallers (MDC_{95}=12.4–15.5 cm/s)[20] and in those poststroke (MDC_{95}=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 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 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 R21AG040387).

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).

REFERENCES

1 Kenny RA, Coen RF, Frewen J, *et al*. Normative values of cognitive and physical function in older adults: findings from the Irish longitudinal study on ageing. *J Am Geriatr Soc* 2013;61:S279–90.
2 Abellan Van Kan G, Rolland Y, Andrieu S, *et al*. Gait speed at usual PACE as a predictor of adverse outcomes in community-dwelling older people an international Academy on nutrition and aging (IANA) Task force. *J Nutr Health Aging* 2009;13:881–9.
3 Cooper R, Kuh D, Hardy R, *et al*. Objectively measured physical capability levels and mortality: systematic review and meta-analysis. *BMJ* 2010;341:c4467.
4 Cooper R, Kuh D, Cooper C, *et al*. Objective measures of physical capability and subsequent health: a systematic review. *Age Ageing* 2011;40:14–23.
5 Beckerman H, Roebroeck ME, Lankhorst GJ, *et al*. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001;10:571–8.
6 Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30:1–15.
7 Glymour MM, Weuve J, Berkman LF, *et al*. When is baseline adjustment useful in analyses of change? an example with education and cognitive change. *Am J Epidemiol* 2005;162:267–78.
8 Rydwik E, Bergland A, Forsén L, *et al*. Investigation into the reliability and validity of the measurement of elderly people's clinical walking speed: a systematic review. *Physiother Theory Pract* 2012;28:238–56.
9 Griswold D, Rockwell K, Killa C, *et al*. Establishing the reliability and concurrent validity of physical performance tests using virtual reality equipment for community-dwelling healthy elders. *Disabil Rehabil* 2015;37:1097–101.
10 Regterschot GRH, Zhang W, Baldus H, *et al*. Test–retest reliability of sensor-based sit-to-stand measures in young and older adults. *Gait Posture* 2014;40:220–4.

11 Ng SS, Hui-Chan CW. The timed up & go test: its reliability and association with lower-limb impairments and locomotor capacities in people with chronic stroke. *Arch Phys Med Rehabil* 2005;86:1641–7.
12 Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc* 1991;39:142–8.
13 Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test. *Phys Ther* 2000;80:896–903.
14 Blankevoort CG, van Heuvelen MJG, Scherder EJA. Reliability of six physical performance tests in older people with dementia. *Phys Ther* 2013;93:69–78.
15 Ries JD, Echternach JL, Nof L, *et al*. Test-retest reliability and minimal detectable change scores for the timed "up & go" test, the six-minute walk test, and gait speed in people with Alzheimer disease. *Phys Ther* 2009;89:569–79.
16 Mangione KK, Craik RL, McCormick AA, *et al*. Detectable changes in physical performance measures in elderly African Americans. *Phys Ther* 2010;90:921–7.
17 Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism. *Phys Ther* 2008;88:733–46.
18 Goldberg A, Chavis M, Watkins J, *et al*. The five-times-sit-to-stand test: validity, reliability and detectable change in older females. *Aging Clin Exp Res* 2012;24:339–44.
19 Hollman JH, Childs KB, McNeil ML, *et al*. Number of strides required for reliable measurements of PACE, rhythm and variability parameters of gait during normal and dual task walking in older individuals. *Gait Posture* 2010;32:23–8.
20 Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables in community-dwelling and hospitalized older fallers. *Gait Posture* 2013;38:1010–4.
21 Brach JS, Perera S, Studenski S, *et al*. The reliability and validity of measures of gait variability in community-dwelling older adults. *Arch Phys Med Rehabil* 2008;89:2293–6.
22 Goldberg A, Schepens S. Measurement error and minimum detectable change in 4-meter gait speed in older adults. *Aging Clin Exp Res* 2011;23:406–12.10.1007/BF03325236
23 Menz HB, Lord SR, St George R, *et al*. Walking stability and sensorimotor function in older people with diabetic peripheral neuropathy. *Arch Phys Med Rehabil* 2004;85:245–52.
24 van Iersel MB, Benraad CEM, Rikkert MGMO. Validity and reliability of quantitative gait analysis in geriatric patients with and without dementia. *J Am Geriatr Soc* 2007;55:632–4.
25 Paterson KL, Hill KD, Lythgo ND, *et al*. The reliability of spatiotemporal gait data for young and older women during continuous overground walking. *Arch Phys Med Rehabil* 2008;89:2360–5.
26 Börsch-Supan A, Brandt M, Hunkler C, *et al*. Data resource profile: the survey of health, ageing and retirement in Europe (share). *Int J Epidemiol* 2013;42:992–1001.
27 UCD Geary Institute and Irish Centre for Social Gerontology N. SHARE Ireland - survey of health, ageing and returement in Europe [Internet], 2008. Available: http://gearyucdie/share/fileadmin/user_upload/shareresults/Share_Wave1_Resultspdf
28 Feeney J, Savva GM, O'Regan C, *et al*. Measurement error, reliability, and minimum detectable change in the Mini-Mental state examination, Montreal cognitive assessment, and color trails test among community living middle-aged and older adults. *J Alzheimers Dis* 2016;53:1107–14.
29 Finucane C, Savva GM, Kenny RA. Reliability of orthostatic beat-to-beat blood pressure tests: implications for population and clinical studies. *Clin Auton Res* 2017;27:31–9.
30 Cronin H, O'Regan C, Finucane C, *et al*. Health and aging: development of the Irish longitudinal study on ageing health assessment. *J Am Geriatr Soc* 2013;61 Suppl 2:S269–78.
31 Guralnik JM, Simonsick EM, Ferrucci L, *et al*. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *J Gerontol* 1994;49:M85–94.
32 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
33 Bohannon RW. Reference values for the five-repetition sit-to-stand test: a descriptive meta-analysis of data from elders. *Percept Mot Skills* 2006;103:215–22.
34 Mehmet H, Yang AWH, Robinson SR. What is the optimal chair stand test protocol for older adults? A systematic review. *Disabil Rehabil* 2019;26:1–8.
35 Lewek MD, Randall EP. Reliability of spatiotemporal asymmetry during overground walking for individuals following chronic stroke. *J Neurol Phys Ther* 2011;35:116–21.