# Test-retest reliability and measurement errors of six mobility tests in the community-dwelling elderly

ORIGINAL ARTICLE

**CY Wang**[1] PhD, **CF Sheu**[2] PhD, **EJ Protas**[3] PhD


## ABSTRACT

**Purpose.** To explore (1) test-retest reliability, (2) measurement error at one point in time (standard error of measurement, SEM), and (3) measurement error between two measurements (smallest real difference, SRD) of 6 mobility tests in the community-dwelling elderly. The SRD can be used to determine whether a real change has occurred.

**Methods.** A physiotherapist administered the 6 mobility tests (functional reach [FR]; usual and fastest gait speed [UGS & FGS]; timed chair stands [TCS]; timed up & go [TUG]; and 6-minute walk distance [6MW]) for 77 community-dwelling elderly in 2 sessions, 1 week apart.

**Results.** Test-retest reliability was excellent (intra-class correlation coefficient$_{2,1}$=0.80-0.95) for all measurements. The Pearson's correlation coefficient (*r*) among the 6 mobility tests varied from 0.2 to 0.9; UGS & FGS, 6MW, and TUG were highly correlated ($r \geq 0.8$). The standard error of the measurement and smallest real difference of all measurements were within 10% and 26%, respectively. The FGS showed the highest test-retest reliability (ICC$_{2,1}$=0.95) and was most responsive for detecting a real change (SRD%=10.3%). The FGS was highly correlated with 6MW (*r*=0.87-0.88) and UGS (*r*=0.79-0.80).

**Conclusion.** The 6 mobility tests provided reliable measurements of mobility functions for the community-dwelling elderly.


**Key words:** Aged; Bias (epidemiology); Geriatric assessment; Reproducibility of results


[1] School of Physical Therapy & Center for Education and Research on Geriatrics and Gerontology, College of Medical Technology, Chung Shan Medical University, Taichung, Taiwan; Department of Physical Therapy, Rehabilitation & Healthy Promotion Center, Chung Shing Hospital, Taichung, Taiwan

[2] Department of Education, National Cheng Kung University, Tainan, Taiwan

[3] Department of Physical Therapy, University of Galveston Medical Branch, Galveston, TX, USA

*Correspondence to: Dr Ching-Yi Wang, School of Physical Therapy, College of Medical Technology, Chung Shan Medical University, No. 110, Sec. 1, Jian-guo N. Rd., South District, Taichung City 40201, Taiwan. E-mail: cywang@csmu.edu.tw*


## INTRODUCTION

Mobility functions such as balance, gait speed, general endurance, and overall mobility are essential for independent living of the community-dwelling elderly. An assessment tool sensitive to real change enables identification of mobility function declines and facilitates implementation of mobility disability prevention programmes for the community-dwelling elderly.

Performance tests to assess mobility functions of the elderly include functional reach (FR), usual and fastest gait speed (UGS & FGS), timed chair stands (TCS), timed up & go (TUG), and 6-minute walk distance (6MW).[1,2] They are useful in monitoring functional change if measurements are stable over time (reliable) and measurement errors are small. There are several studies on the test-retest reliability of these mobility tests in the community-dwelling elderly.[3-7] None reported the measurement errors for determining real change (i.e. whether the change is beyond measurement error). In addition, as different mobility tests were used in different studies, comparisons are difficult.

Descriptions of mobility status (i.e. healthier, normal, or less healthy) and its change (i.e. improving,

no change, or deteriorating) are clinically useful. To provide meaningful information to the health care providers (i.e. whether a status/score truly changes), the standard error of the measurement (SEM) and variability between test and retest at 95% confidence interval (CI)—smallest real difference (SRD)—should be provided.[8-11] The measurement error is also important for calculating statistical power and estimating sample sizes.[11]

This study aimed to assess (1) test-retest reliability, (2) measurement error at one point in time (SEM), and (3) measurement error between 2 measurements (SRD) at 95% CI of 6 mobility tests in the community-dwelling elderly.

## METHODS

### Participants

Mobility functions of 47 men and 30 women (mean age, 73±6 years) with a mean body mass index of 24±3 kg/m$^2$ were assessed. Inclusion criteria were (1) age 65 years or older, (2) living in the community, and (3) able to perform all 6 tests independently without using any assistive device. Those who had unstable high blood pressure or a history of a serious health problem, fell ill, were in pain, or had participated in any activity or training programme in the past week that might affect their performance were excluded.

### Procedure

All subjects gave informed consent prior to testing. They were asked if they had eye sight problems, hearing problems, dental problems, hypertension, diabetes, heart disease, arthritis, or other health problems. Participants self-rated their health status as 'healthier', 'the same', or 'less healthy' compared to others of their age. They also self-reported their mobility (ability to walk several blocks, to walk up and down stairs, and to perform heavy housework),[12] instrumented activities of daily living (IADL) [preparing meals, shopping, doing light housework, taking medication, using transportation, using the telephone, handling finances],[13] and activities of daily living (ADL) [eating, dressing, bathing, walking across a small room, transferring from bed to chair, grooming, and using toilet].[14] A domain was rated as 'disabled' when the participant reported 'needing help' or 'unable to perform' on any single item in that domain. Their mental status was assessed with the Chinese version of Mini-Mental Status Examination.[15]

All subjects were tested at a local community centre on 2 occasions in the same way and at the same time of a day approximately 1 week apart (mean, 8.1±2.5 days). A single physiotherapist blinded to the purpose of this study administered all tests. In each session, the 6MW was performed once and at the end to avoid fatigue on other tests. The other 5 mobility tests were performed twice. To make it safer and more comfortable for the participants, the UGS always preceded the FGS, whereas the testing order for the FR, TUG, and TCS was randomised. Prior to each test, the physiotherapist demonstrated the test to the participant. Subjects were allowed to rest as long as they wanted before proceeding to the next test.

### Mobility tests

The FR measured the distance a subject could reach forward without taking a step when standing beside a wall with the shoulder adjacent to a yardstick, the arm in 90º of flexion, and the hand in a fist.[16] The distance (cm) between the starting position and the farthest reaching position was recorded.

UGS and FGS were calculated as m/s for the distance of 15.24 m.[17] The subjects stood at the start line and walked past the end point at their usual/fastest speed. Timing began when one foot passed the start point and stopped when one foot passed the end point. Subjects were allowed a short rest in between tests.

TUG measured the time (in seconds) required to stand up from a chair (45 cm in height), walk 3 meters, turn, walk back to the chair, turn, and sit down.[18] The subjects first sat on the chair and were cued to start using "one, two, go" and performed the manoeuvre at their fastest but safest speed. Shorter times indicated better mobility.

TCS measured the time (in seconds) for completing 5 consecutive chair stands as quickly as possible.[19] Subjects sat on the chair with their feet flat on the floor and stood up without using their arms. Timing began when the tester said "go" and ended when the subjects sat down on the chair for the fifth time.

**Table 1**
**Means, standard deviations, 95% CI of mean, and relative retest reliability ($ICC_{2,1}$) of the 6 mobility tests on 2 sessions**

| Mobility tests | Mean±SD | | $ICC_{2,1}$ (95% CI) |
|---|---|---|---|
| | Session 1 | Session 2 | |
| Timed chair stands (s) | 9.9±2.6 | 9.9±2.7 | 0.89 (0.83-0.93) |
| Functional reach (cm) | 29.4±4.1 | 29.9±4.3 | 0.84 (0.75-0.89) |
| Usual gait speed (m/s) | 1.3±0.2 | 1.3±0.2 | 0.80 (0.70-0.87) |
| Fastest gait speed (m/s) | 1.6±0.3 | 1.6±0.3 | 0.95 (0.92-0.97) |
| Timed up & go (s) | 7.4±1.4 | 7.3±1.5 | 0.90 (0.85-0.94) |
| 6-minute walk distance (m) | 496.1±92.2 | 507.4±90.2 | 0.92 (0.87-0.95) |

6MW measured the maximum distance the subjects could walk in 6 minutes in a 12-meter long square course, free of obstacles, in the community centre. Subjects were allowed to slow down, stop, and rest. Standardised verbal encouragement was given every minute, such as"you are doing well; you have x minute(s) to go".

### Data analysis

The means and standard deviations of the mobility tests for each session were reported. The associations among the 6 mobility tests during each session were analysed by the Pearson's correlation coefficients at a 0.05 significance level.

The relative retest reliability of the mobility tests was determined by the intra-class correlation coefficient ($ICC_{2,1}$) using a 2-way random-effects ANOVA model (subject by session) and the absolute agreement definition.[11,20,21] The 95% CIs for the ICCs were calculated. Values of retest reliability of a measurement above 0.8 indicated good reliability.[22,23] The systematic error (the mean of difference scores of retest and test) was checked by a paired *t*-test with the significance level set at 0.05.

The absolute retest reliability was assessed by the SEM and SRD.[11,24] The SEM represented the variability between measurements obtained from the 2 sessions and was calculated as the square root of the within subjects error variance (SEM=√WMS).[17,25,26] The SRD demonstrated the 95% CI of the difference in score between paired observations and was calculated as 1.96×√2×SEM.[8,25,27]

The SEM% (SEM/mean of all measurements from both sessions x100) and the SRD% (SRD/mean of all test-retest measurements x100) were also reported for comparison among the 6 tests using different units of measurement.[25]

## RESULTS

Of the 77 participants, 48 (62%) reported no disability, 23 (30%) reported mobility disability only, and 6 (8%) reported both mobility and IADL limitations. None had restrictions in all 3 domains. 46% of the participants perceived their health as healthier, 35% as the same, and 20% as less healthy than others of a similar age. 51% reported at least 2 comorbidities, the most common being eyesight problems (60%), high blood pressure (46%), and arthritis (21%). The mean Mini-Mental Status Examination score was 26.4±2.5.

Regarding the Pearson's correlation coefficients among the 6 mobility tests, the strongest was between the 6MW and FGS ($r$=0.87-0.88, $p<0.05$), followed by the FGS and the UGS ($r$=0.79-0.80, $p<0.05$), and the 6MW and TUG ($r$=-0.76 to -0.79, $p<0.05$). The correlations between other pairs of mobility tests varied from 0.19 to 0.75 (**Table 1**).

The retest reliabilities ($ICC_{2,1}$) of all measurements varied from 0.80 to 0.95 (**Table 1**). The FGS showed the highest retest reliability ($ICC_{2,1}$=0.95), followed by the 6MW ($ICC_{2,1}$=0.92) and the TUG ($ICC_{2,1}$=0.90). The UGS was lowest ($ICC_{2,1}$=0.80). The test-retest reliabilities of all 6 mobility tests exceeded 0.8. The paired *t*-tests showed non-significant differences between the means of the 2 sessions across all measurements.

The SEM% of the 6 mobility tests were all within 10%, the lowest being for the FGS (3.7%) and the

**Table 2**
**Absolute reliability indices of the 6 mobility tests**

| Mobility tests | Absolute reliability indices* | | | |
|---|---|---|---|---|
| | SEM | SEM% | SRD | SRD% |
| Timed chair stands (s) | 0.9 | 9.1 | 2.5 | 25.0 |
| Functional reach (cm) | 1.7 | 5.7 | 4.7 | 15.8 |
| Usual gait speed (m/s) | 0.09 | 7.0 | 0.25 | 19.3 |
| Fast gait speed (m/s) | 0.06 | 3.7 | 0.17 | 10.3 |
| Timed up & go (s) | 0.4 | 5.5 | 1.1 | 15.0 |
| 6-minute walk distance (m) | 24.7 | 4.9 | 68.4 | 13.6 |

* SEM denotes standard error of measurement, and SRD smallest real differences

highest for the TCS (9.1%). The FGS demonstrated the smallest SRD% (10.3%), followed by the 6MW (13.6%) and the TUG (15%) [**Table 2**].

## DISCUSSION

Among the 6 mobility tests, FGS showed the highest test-retest reliability ($ICC_{2,1}$=0.95) and was most responsive for detecting a real change (SRD%=10.3%). Test-retest reliability of these mobility tests in the community-dwelling elderly was good to excellent, which was consistent with previous studies.[4,5,28,29] In a study on 10 healthy community-dwelling elderly, the 1-week test-retest reliability of the TUG was excellent ($ICC_{3,1}$=0.97) and the 1-week retest reliability of the 6MW was ICC=0.91.[29] In another study, the 2-week test-retest reliabilities of the FR and TUG were excellent (ICC=0.93-0.99).[4] The test-retest reliability of the 6MW over 3 trials with an interval of 3 to 5 days was high between trials 1 and 2 (ICC=0.91) and between trials 2 and 3 (ICC=0.94).[28] In a study using a 4-m distance to assess the gait speed, the test-retest reliability of UGS was excellent (ICC=0.94).[5] In our study using a 15.2-m distance, the 1-week retest reliability was good for UGS ($ICC_{2,1}$=0.80) and excellent for FGS ($ICC_{2,1}$=0.95), which was consistent with findings on stroke patients.[17] However, in the frail, heterogeneous community-dwelling elderly, the retest reliability of these mobility tests was moderate (ICC=0.64-0.79),[3,6,7] indicating that the retest reliability of a measurement needs to be determined for the population of interest.

The SEM quantifies the measurement error of an observation at one point in time, in reference to the precision of individual scores on a test, and has been used to define the boundaries around which a subject's true score lies.[23] Thus, it can be used to calculate a range where the true score of a community-dwelling elderly is located at 95% CI (±1.96×SEM, 1.96 is the *z* score associated with a 95% CI).

A change in mobility status between test and retest results (i.e. improving, no change or deteriorating) is based on the change in score and its error size (SRD). When a change in score (retest-test) is greater than the SRD, a true change can be ascertained with a 95% CI. For example, the SRD for the 6MW was 68.4 meters. If a change in score between retest and test is greater than +68.4 (or -68.4) meters, then a true improvement (or deterioration) is deemed to have occurred with a 95% CI. It is easier to use SRD (random error of the change in score) as a criterion to determine whether a real change has occurred.[11]

Retest reliability, measurement errors, and sensitivity to real change of the TUG, UGS, FGS, and 6MW differ between the community-dwelling elderly and stroke patients with hemiparesis.[17] The mobility of stroke patients is significantly lower than our community-dwelling elderly (their means were not within the 95% CI of our mean).[17] The retest reliability of the 6MW (but not the TUG, UGS, and FGS) differed significantly in both populations (the 95% CI of retest reliability of both populations did not overlap). Compared to post-stroke patients, our community-dwelling elderly showed larger SEM in UGS and 6MW scores, and smaller SEMs for the TUG and FGS. The tests with the smallest measurement errors and most sensitive to real change were 6MW (SEM%=4.8%, SRD%=13%) for post-stroke patients and FGS (SEM%=3.7%,

SRD%=10.3%) for the community-dwelling elderly. The tests with the largest measurement errors and least responsiveness to real change were TUG (SEM%=8.2%, SRD%=23%) for post-stroke patients and UGS (SEM%=7.0%, SRD%=19.3%) for the community-dwelling elderly. The SEM% and SRD% provided similar information that less measurement error indicated more sensitivity to real change.

The SRD differs from the 'clinically important change'.[9,30] A test responsive to a 'real change' does not imply that it is also responsive to a 'clinically important change'. Nonetheless, the SRD is crucial in determining whether the magnitude of the effect is clinically important. A change must be larger than the size of measurement error before being considered important or meaningful.[9,10] Yet in situations when the 'clinically important change' is smaller than the SRD, other measurements with an SRD smaller than the 'clinically important change' (responsiveness to important change) should be identified to quantify this important change.

### Limitations

Our non-randomised sample was relatively medically stable. 30% of our participants reported an early stage of physical disability (needing help or inability to perform at least one task in the mobility domain) and 8% reported disability in both mobility and IADL domains. Thus, our results can be used for monitoring the changes of mobility status in the community-dwelling elderly. Generalisation of these results is limited to the population studied and may not be appropriate for frail or institutionalised persons. As the reliability of a test is dependent on the characteristics of the sample and the conditions under which the measurement is made, the measurement error for determining real change for individual or group comparisons must be reported for frail persons or those receiving interventions. This study was based on a 1-week retest interval. The retest measurement error may change if the test-retest interval is not 1 week. Thus, there may be incorrect interpretation of changes in performance if our results are applied to measurements taken at different intervals. The influence of the test-retest interval on the magnitude of measurement error should be determined in future researches. A study of serial retests over a longer period may yield more stable estimates of error. The 6 mobility tests have been widely used in assessing mobility functions of the elderly. Other mobility tests that can predict/discriminate important health outcomes such as falls, disability, or death, should also be examined in future studies.

## CONCLUSION

The 6 mobility tests provided reliable measurements of mobility functions for the community-dwelling elderly. The SRD could determine whether a real change occurred at 95% CI. Among the 6 mobility tests, FGS showed the highest test-retest reliability ($ICC_{2,1}$=0.95) and was the most responsive for detecting a real change (SRD%=10.3%). FGS also showed strong correlation with 6MW ($r$=0.9, $p<0.05$) and UGS ($r$=0.8, $p<0.05$), but weak-to-moderate correlation with other tests.

## ACKNOWLEDGEMENT

This study was partially supported by a grant from National Science Council of Taiwan (NSC 94-2314-B-277-004).

### References

1. Steffen TM, Hacker TA, Mollinger L. Age- and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds. *Phys Ther* 2002;82:128-37.
2. Lusardi MM, Pellecchia GL, Schulman M. Functional performance in community living older adults. *J Geriatr Phys Ther* 2003;26:14-22.
3. Jette AM, Jette DU, Ng J, Plotkin DJ, Bach MA. Are performance-based measures sufficiently reliable for use in multicenter trials? Musculoskeletal Impairment (MSI) Study Group. *J Gerontol A Biol Sci Med Sci* 1999;54:M3-6.
4. Lin MR, Hwang HF, Hu MH, Wu HD, Wang YW, Huang FC. Psychometric comparisons of the timed up and go, one-leg stand, functional reach, and Tinetti balance measures in community-dwelling older people. *J Am Geriatr Soc* 2004;52:1343-8.
5. Rolland YM, Cesari M, Miller ME, Penninx BW, Atkinson HH, Pahor M. Reliability of the 400-m usual-pace walk test as an assessment of mobility limitation in older adults. *J Am Geriatr Soc* 2004;52:972-6.
6. Ostchega Y, Harris TB, Hirsch R, Parsons VL, Kington R, Katzoff M. Reliability and prevalence of physical performance examination assessing mobility and balance in older persons in the US: data from the Third National Health and Nutrition Examination Survey. *J Am Geriatr Soc* 2000;48:1136-41.
7. Rockwood K, Awalt E, Carver D, MacKnight C. Feasibility and measurement properties of the functional reach and the timed up and go tests in the Canadian study of health and aging. *J Gerontol A Biol Sci Med Sci* 2000;55:M70-3.
8. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001;10:571-8.
9. Finch E, Brooks D, Stratford PW, Mayo NE. Physical rehabilitation

outcome measures. *A guide to enhanced clinical decision making*. 2nd ed. Canada: Canadian Physiotherapy Association; 2002:26-41.
10. Schuck P, Zwingmann C. The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. *Int J Rehabil Res* 2003;26:85-91.
11. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217-38.
12. Rosow I, Breslau N. A Guttman health scale for the aged. *J Gerontol* 1966;21:556-9.
13. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. *Gerontologist* 1969;9:179-86.
14. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychological function. *JAMA* 1963;185:914-9.
15. Chiu HF, Lee HC, Chung WS. Reliability and validity of the Cantonese version of Mini-Mental Status Examination: A preliminary study. *J Hong Kong Coll Psychiatry* 1994;4:25-8.
16. Duncan PW, Weiner DK, Chandler J, Studenski S. Functional reach: a new clinical measure of balance. *J Gerontol* 1990;45: M192-7.
17. Flansbjer UB, Holmback AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med* 2005;37:75-82.
18. Podsiadlo D, Richardson S. The timed "Up and Go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc* 1991;39:142-8.
19. Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. *N Engl J Med* 1995;332:556-61.
20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.
21. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30-46.
22. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskelet Disord* 2005;6:3.
23. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231-40.
24. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135-60.
25. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. *Am J Phys Med Rehabil* 2005;84:719-23.
26. Schreuders TA, Roebroeck ME, Goumans J, van Nieuwenhuijzen JF, Stijnen TH, Stam HJ. Measurement error in grip and pinch force measurements in patients with hand injuries. *Phys Ther* 2003;83:806-15.
27. Lim LI, van Wegen EE, de Goede CJ, Jones D, Rochester L, Hetherington V, et al. Measuring gait and gait-related activities in Parkinson's patients own home environment: a reliability, responsiveness and feasibility study. *Parkinsonism Relat Disord* 2005;11:19-24.
28. Rikli RE, Jones J. The reliability and validity of a 6-minute walk test as a measure of physical endurance in older adults. *J Aging Phys Act* 1998;6:363-75.
29. 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.
30. Hebert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. *Arch Phys Med Rehabil* 1997;78:1305-8.