Background and Purpose. It is unknown how cognitive impairment affects the reliability of Timed “Up & Go” Test (TUG) scores. The aim of the present study was to investigate the expected variability of TUG scores in older subjects dependent in activities of daily living (ADL) and with different levels of cognitive state. The hypothesis was that cognitive impairment would increase the variability of TUG scores.

Subjects. Seventy-eight subjects with multiple impairments, dependent in ADL, and living in residential care facilities were included in this study. The subjects were 84.8±5.7 (mean±SD) years of age, and their Mini-Mental State Examination score was 18.7±5.6.

Methods. The TUG assessments were performed on 3 different days. Intrarater and interrater analyses were carried out.

Results. Cognitive impairment was not related to the size of the variability of TUG scores. There was a significant relationship between the variability and the time taken to perform the TUG. The intraclass correlations were greater than .90 and were similar within and between raters. In repeated measurements at the individual level, an observed value of 10 seconds was expected to vary from 7 to 15 seconds and an observed value of 40 seconds was expected to vary from 26 to 61 seconds for 95% of the observations.

Discussion and Conclusion. The measurement error of a TUG assessment is substantial for a frail older person dependent in ADL, regardless of the level of cognitive function, when verbal cuing is permitted during testing. The variability increases with the time to perform the TUG. Despite high intraclass correlation coefficients, the ranges of expected variability can be wide and are similar within and between raters. Physical therapists should be aware of this variability before they interpret the TUG score for a particular individual.


Key Words: Cognition, Frailty, Geriatric assessment, Mobility, Reliability, Timed “Up & Go” Test, Variability.

Ellinor Nordin, Erik Rosendahl, Lillemor Lundin-Olsson
Various assessments of physical mobility and movement strategies are used in geriatric rehabilitation. Whether the purpose is screening, detecting a change in ability, or making predictions, methods that yield valid and reliable data are essential. In addition, when older people with cognitive and physical disabilities are assessed, it is also particularly important for the methods used in measurement to be applied easily.

The Timed “Up & Go” Test (TUG)\(^1\) is a simple and inexpensive method that was developed to screen basic mobility. The TUG comprises basic everyday movements: stand up from a chair, walk 3 m, turn around, walk back, and sit down again. The outcome is the time taken to perform this sequence of movements. Within research, the use of the TUG has increased over the last few years, and it is recommended by the British Geriatrics Society, the American Geriatrics Society, and Nordic geriatricians when screening for risk of falling.\(^2,3\) The TUG is reported to yield reliable and valid data for use in older people.\(^1,4–12\) However, studies that have evaluated the reliability and validity of TUG scores have included mainly medically stable subjects with good or only mildly affected cognitive functions. To our knowledge, only 2 studies\(^13,14\) have attempted to evaluate the effect of cognitive dysfunction on the performance of the TUG. Rockwood et al\(^13\) found poor test-retest reliability; however, because the time interval between the measurements was 112±72 (mean±SD) days, this issue cannot be said to have been explored correctly. Thomas and Hageman\(^14\) found excellent test-retest reliability; however, their sample of 9 subjects seems to be too small to draw conclusions regarding reliability. Therefore, it is still not known how the presence of cognitive impairment affects the reliability of TUG scores.

There is some support for the notion that cognitive impairment could have a negative influence on the reliability of geriatric assessments in older nursing home residents.\(^15\) There is a known correlation between cognitive function and gait. Slower walking speed, decreased step length, increased double-support ratio, and step-to-step variability have been found more often in people with cognitive impairment than in older adults who are healthy.\(^16,17\) This finding could be explained by a deterioration of cerebral integration and processing of sensory (visual, vestibular, and proprioceptive) information.\(^16,17\) Accordingly, cognitive impairment may result in increased variability in repeated measures of TUG assessment, as this test requires an understanding of the instructions as well as an interaction among the patient, the assessor, and the environmental setting.

There is always some variability in repeated measurements; this variability depends on the person being measured, on the assessor who does the measuring, and on the precision of the instrument.\(^18,19\) This variability can represent systematic bias (learning or fatigue effects on the test) or random error attributable to inherent biological or mechanical variations. Reliability can be considered to be the amount of measurement error that has been judged to be acceptable for the valuable use of a measurement tool,\(^20\) and the determination of reliability is a clinical decision based on evaluations that include statistical analyses. Consequently, it is important that the

E Nordin, PT, MSc, is a doctoral student, Department of Community Medicine and Rehabilitation, Physiotherapy and Geriatric Medicine, Umeå University, SE-901 87 Umeå, Sweden (ellinor.nordin@physiother.umu.se). Address all correspondence to Mrs Nordin.

E Rosendahl, PT, MSc, is a doctoral student, Department of Community Medicine and Rehabilitation, Physiotherapy and Geriatric Medicine, Umeå University.

L Lundin-Olsson, PT, PhD, is Researcher, Department of Community Medicine and Rehabilitation, Physiotherapy and Geriatric Medicine, Umeå University.

All authors provided concept/idea/research design data collection and analysis, project management, and facilities/equipment. Ms Nordin provided writing. Dr Lundin-Olsson provided subjects and institutional liaisons. Mr Rosendahl and Dr Lundin-Olsson provided consultation (including review of manuscript before submission). The authors thank their colleagues who participated in the testing procedure for their valuable contribution to this study.

This study was approved by the Ethics Committee of the Medical Faculty of Umeå University.

This investigation was supported by grants from the Aldrecentrum Vasterbotten, the Minnesfond of the Registered Physiotherapy Association, the Erik and Anne-Marie Detlof’s Foundation of Umeå University, the Swedish Council for Working Life and Social Research, and the Osterman Research Foundation.

This research was presented as a poster presentation at the International Congress of The World Confederation for Physical Therapy, June 7–12, 2003, Barcelona, Spain, and as an oral presentation at a national conference for physical therapists arranged by the Swedish Association of Registered Physiotherapists, October 15–17, 2003, Stockholm, Sweden.

This article was received March 21, 2005, and was accepted December 7, 2005.
The aim of the present study was to investigate the expected variability of TUG scores in older subjects dependent in ADL and with different levels of cognitive state. The hypothesis was that cognitive impairment would increase the variability of TUG scores.

**Method**

**Settings and Participants**

In Sweden, older people living in residential care facilities are disabled as a result of cognitive or physical impairments or both; in addition, they require supervision, functional support, or nursing care. Participants in this study were people aged 65 years or older and living in 9 different residential care facilities located in Umeå, a city in northern Sweden. Of the 487 screened subjects, 78 met the inclusion criteria: Mini-Mental State Examination (MMSE) scores of ≥10; dependence in 1 or more personal ADL, according to the Katz index; and eligibility and ability to perform the TUG without personal physical assistance on 3 occasions within 7 days.

The characteristics of the 78 participants are shown in Table 1. Sixty-two percent were women, and the mean age was 84.8 years. The mean MMSE score was 18.7 points. As indicated by the suggested classifications, 19 participants had no cognitive impairment, 23 had mild impairment, and 36 had severe impairment. According to the Barthel Index, 72 of them were dependent when taking a shower or bath, 38 of them were totally or partly dependent when getting dressed, and 7 required assistance when eating. In addition, there were wide ranges of balance abilities and of walking speeds. The great

---

**Table 1.**

Characteristics of the Study Participants (N=78)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women, no. (%)</td>
<td>49 (62)</td>
</tr>
<tr>
<td>Age, y, mean±SD (range)</td>
<td>84.8±5.7 (66–97)</td>
</tr>
<tr>
<td>Diagnosis, no. (%) of participants</td>
<td></td>
</tr>
<tr>
<td>Dementia</td>
<td>41 (53)</td>
</tr>
<tr>
<td>Depression</td>
<td>38 (49)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>23 (30)</td>
</tr>
<tr>
<td>Delirium (the last year)</td>
<td>20 (26)</td>
</tr>
<tr>
<td>Heart failure</td>
<td>18 (23)</td>
</tr>
<tr>
<td>Angina pectoris</td>
<td>17 (22)</td>
</tr>
<tr>
<td>Osteoarthrosis</td>
<td>14 (18)</td>
</tr>
<tr>
<td>Stroke</td>
<td>13 (17)</td>
</tr>
<tr>
<td>Chronic lung disease</td>
<td>7 (9)</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>1 (1)</td>
</tr>
<tr>
<td>No. of prescribed drugs/d, mean±SD (range)</td>
<td>8.6±4.8 (0–27)</td>
</tr>
<tr>
<td>Cognitive and physical state</td>
<td></td>
</tr>
<tr>
<td>MMSE score, mean±SD (range)</td>
<td>18.7±5.6 (10–29)</td>
</tr>
<tr>
<td>Barthel Index score, mean±SD (range)</td>
<td>14.9±3.0 (5–19)</td>
</tr>
<tr>
<td>Berg Balance Scale score, mean±SD (range)</td>
<td>33.2±11.8 (8–55)</td>
</tr>
<tr>
<td>Self-paced walking speed, m/s, mean±SD (range)</td>
<td>0.4±0.2 (0.2–0.8)</td>
</tr>
<tr>
<td>Use of walking aid indoors, no. (%) of participants</td>
<td>43 (55)</td>
</tr>
<tr>
<td>Independent gait indoors, with or without aid, no. (%) of participants</td>
<td>64 (82)</td>
</tr>
</tbody>
</table>

* MMSE = Mini-Mental State Examination.

* Data were missing for 6 subjects.
majority of participants were independent walkers indoors, and more than half normally used a walking aid.

Assessments
Assessments were made by 12 physical therapists working within the study and 8 physical therapists employed by the local authority with responsibilities for the different facilities. The methodological framework was considered in order to create optimal conditions for the evaluation. Before baseline assessments, a physical therapist with long experience using the TUG in older people (EN) gave all physical therapists involved in the testing procedures both verbal and written instructions based on the original description of the TUG assessment,1 complementary directives for the start and stop of timing,5 and instructions to record possible verbal cuing. All physical therapists performed trial assessments. This training was done in 1 session for the study physical therapists and in another session for the physical therapists employed by the local authority. The physical therapists also were introduced to and practiced administering the other instruments used in the present study to describe a participant’s state: MMSE to estimate the participant’s cognitive state,21 the Berg Balance Scale,24 a 2.4-m self-paced walk to assess balance and ambulation,25,26 and the Barthel Index27 to screen for dependence in ADL.

The MMSE21 is a screening tool that includes 11 questions in 6 sections, each representing a different cognitive domain or function (orientation, registration, attention and calculation, recall, language, and copying). The maximum score is 30. A score of 23 points or less has been considered to be evidence of cognitive impairment, scores between 18 and 23 points indicate mild impairment, and scores of 17 or less indicate severe impairment.23,28 We chose a cutoff for MMSE scores of ≥10 for inclusion in the study. Our reason for using this criterion was based on our clinical experience that people with very low MMSE scores have difficulties in following instructions. This reasoning also was in accordance with the findings of Mozley et al,29 who reported this level to be important for the ability to communicate and interact in a meaningful way.

In order to present a more complete clinical description and understanding of the studied sample, the functional capacity of the subjects was assessed by use of the Berg Balance Scale, analysis of gait speed, and the Barthel Index of ADL, and their medical state was assessed by recording of diagnosis and prescribed drugs. The Berg Balance Scale is a measure of balance; it contains 14 different items, including sitting, standing with different bases of support or with eyes closed, turning, and picking an object up from the floor.24 The maximum score is 56, indicating good balance abilities for an older person. Gait speed (in meters per second) was measured by use of a stopwatch and a 2.4-m walk at a speed comfortable for the participants.25,26 Participants were instructed to start from standing, using their ordinary walking aid, and to walk to a goal approximately 4 to 5 m away. The distance of 2.4 m was marked discretely with tape on the wall. In addition, the physical therapists interviewed the staff at the facilities with regard to the residents’ ADL according to the Barthel Index. This instrument measures the degree of independence in 3 categories of everyday living: self-care (washing and showering, dressing, eating, and toilet visits), continence of bowel and bladder, and mobility (transfer to and from a chair, toilet, and a tub or shower; walking indoors; and climbing stairs). The maximum score of 20 indicates independence in personal ADL.27 A registered nurse at each facility completed a questionnaire regarding diagnosis and prescribed drugs. This questionnaire was reviewed by a specialist in geriatric medicine before completion of the diagnosis.

The residents participated in 3 test sessions of TUG on different days, all within 1 week: TUG A (day 1), TUG B (day 2), and TUG C (day 3). To avoid any change that might normally occur in a frail subject over a day, for example, as a result of medication or food intake, each session of TUG assessment was carried out at approximately the same time of day as the preceding session for that subject. The study physical therapists assessed TUG A and TUG B (intrarater reliability), and the local authority physical therapists assessed TUG C (intrarater reliability). With this procedure, multiple paired observations were made.

In accordance with the directions given by Podsiadlo and Richardson,1 the TUG was performed twice in each test session, that is, 1 trial and 1 timed performance, with only a brief seated rest in between. The physical therapist first demonstrated how to perform the TUG and then instructed the participants to stand up from a chair, walk 3 m at a comfortable speed, cross a line on the floor, turn around, walk back, and sit down again. An armchair of standard height, approximately 45 cm, was used. The starting position was sitting with the back against the chair back, with arms resting on the arm support of the chair, and with the walking aid near at hand. Participants wore their usual indoor footwear and chose whether to use their walking aid when performing the TUG. No personal physical assistance was allowed. The physical therapist either was standing next to the chair or accompanied the participant if required for reasons of safety, such as in the case of gait difficulties. In the case of a slight hesitation from the participant on how to proceed with the TUG, the physical therapist gave additional verbal cuing, such as “turn around” or...
“sit down.” The use of verbal cuing during the test was recorded in the test protocol as present or not present. The TUG was timed with an ordinary stopwatch. The start and stop of timing were fixed firmly. Timing started when the participant’s back was no longer in contact with the back of the chair. Timing stopped when their buttocks touched the seat of the chair again. The result, in seconds to 1 decimal place, was recorded.

The results of different test sessions were recorded on separate documents. Physical therapists were unaware of the previous results. The information transferred was limited to date and time of the previous test session as well as what kind of walking aid, if any, was used in TUG A. The same walking aid was used for all test sessions. The residents were given written and oral information about the study. The participants or their proxies, when appropriate because of cognitive impairment, gave their consent for participation.

Data Analysis

The differences in times for TUGs between test sessions were calculated by subtracting the TUG B score from the TUG A score (intrarater reliability) and by subtracting the TUG C score from the TUG A score (interrater reliability). These calculations would result in a negative time difference if the second performance (TUG B or TUG C) were slower than performance on TUG A and a positive difference if it were faster. The distribution of TUGs and the difference in times between TUGs were skewed; therefore, the data were transformed logarithmically before any further analyses.

Linear regression analysis was used to evaluate the effect of cognition on the difference in times between TUGs. The dependent variables were the difference between TUG A and TUG B for intrarater analyses and the difference between TUG A and TUG C for interrater analyses. The independent variables were the MMSE score and the MMSE score dichotomized (cutoffs between 17 and 18 and between 23 and 24). Analyses were done with and without adjustments for TUG A. The rationale behind the adjustment was that the time needed to perform the TUG was considered to be a possible confounder for the difference between 2 TUG performances.

Logistic regression analysis was used to evaluate the effect of cognitive function in the presence of verbal cuing given during the performance of the TUG. The dependent variable was verbal cuing (yes or no), and the independent variable was the MMSE score. Analyses were done with and without adjustments for the TUG score because the time needed to perform the TUG was considered to be a possible confounder for verbal cuing.

The reliability of TUG scores was analyzed by use of TUG A and TUG B for analyses within raters and TUG A and TUG C for analyses between raters. We chose to present the results of both relative reliability and absolute reliability.18–20 The relative reliability was analyzed by use of 2 different models of the ICC with a 95% confidence interval (CI).30 With ICC(1,1), it is assumed that all error is random measurement error. With ICC(3,1), it is assumed that systematic error is not part of the measurement error. No systematic error is present when ICC(1,1) equals ICC(3,1).31 We calculated ICC(1,1) with a 1-way random-effects model for a single measure, and ICC(3,1) was calculated with a 2-way mixed model for a single measure. The absolute reliability was analyzed by use of the within-subject standard deviation (S_w) as described by Bland and Altman.32,33 The first step was to verify the distribution of the data by plotting the difference between 2 TUG measures against their mean. We found a relationship between the measurement error and the magnitude of the TUG scores; that is, heteroscedasticity was present (Fig. 1). In such a case, log-transformed data are used in the analyses.20,32,33 The S_w then was obtained from the square root of the within-

Figure 1.
Differences in Timed “Up & Go” Test (TUG) results between TUG A and TUG B plotted against the mean for TUG A and TUG B (in seconds) on the natural scale for each subject. The horizontal line indicates equality between measurements.
The Statistical Package for the Social Sciences (SPSS), version 11.0,* was used for statistical analyses. The level of significance was set to $P<.05$.

Results

The TUG A, TUG B, and TUG C assessments were made within 5 days in 83% of the cases, and the remainder were made within 7 days. On average, there was a 1-hour difference between the time of day when TUG B and TUG C were carried out and the time of day when TUG A was carried out.

The time needed to perform the TUG ranged from 9.8 to 109.2 seconds. The distribution was skewed, with a mean and a median of approximately 30 and 25 seconds, respectively. The results are shown in Table 2.

Table 2.
Results of the Timed "Up & Go" Test (TUG) From 3 Different Days Within 1 Week [N=78]

<table>
<thead>
<tr>
<th>Parameter (Score Range)</th>
<th>TUG Score, s (Interquartile Range, Q1–Q3)</th>
<th>Mean±SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUG d 1 [11.3–87.7]</td>
<td>24.1 (17.9 to 37.1)</td>
<td>30.0±17.4</td>
</tr>
<tr>
<td>TUG d 2 [11.3–109.2]</td>
<td>24.4 (16.7 to 36.0)</td>
<td>29.9±17.7</td>
</tr>
<tr>
<td>TUG d 3 [9.8–83.4]</td>
<td>25.0 (16.6 to 35.4)</td>
<td>28.0±15.1</td>
</tr>
<tr>
<td>Difference between d 1 and d 2\a</td>
<td>0.2 (–2.1 to 3.2)</td>
<td>0.1±9.0</td>
</tr>
<tr>
<td>Difference between d 1 and d 3\b</td>
<td>0.6 (–3.1 to 3.4)</td>
<td>2.0±8.5</td>
</tr>
</tbody>
</table>

* Same rater.
\a Different raters.

Effect of Cognition

Analyses of the differences in times between TUGs were carried out with and without adjustments for the time needed to perform TUG. The MMSE score was not related to the difference between TUG A and TUG B (unadjusted $P=.480$, adjusted $P=.903$) or to the difference between TUG A and TUG C (unadjusted $P=.250$, adjusted $P=.730$). Analyses with dichotomized MMSE scores to divide participants into subgroups did not alter these findings. The adjusted analyses revealed a significant relationship between slower TUG performance and the difference between TUG A and TUG B as well as the difference between TUG A and TUG C ($P<.001$).

Verbal cuing during testing was given to 77% of the participants in TUG A, to 68% in TUG B, and to 60% in TUG C. In the analyses without adjustment for possible confounding by the time needed to perform TUG, the MMSE scores were significantly related to participants who were cued in TUG B and TUG C ($P=.038$ and $P=.002$, respectively) but not in TUG A ($P=.101$). When the analyses were adjusted for the time needed to perform the TUG, it was shown that a slow performer was more likely to be cued and that the cognitive state was of importance only on day 3 (day 1: TUG A $P=.010$, MMSE $P=.217$; day 2: TUG B $P=.055$, MMSE $P=.090$; day 3: TUG C $P=.023$, MMSE $P=.006$).

Reiability

The median and interquartile range (Q1–Q3) of the difference between test sessions reached ICC(1,1) values of .86 for intrarater measurements, respectively; these values were 0.6 and –3.1 to 3.4 seconds for interrater measurements, respectively (Tab. 2). Relative reliability between test sessions reached ICC(1,1) values of .92 (95% CI=.86–.95) for intrarater measurements and .91 (95% CI=.86–.94) for interrater measurements. The corresponding values for ICC(3,1) were .91 (95% CI=.87–.94) and .91 (95% CI=.86–.94). Thus, no systematic error was present. Figure 1 shows a plot of the mean difference (in seconds) between TUG A scores and TUG B scores before log transformation.

The antilog$S_w^{1.96}$ within raters was 1.3507, and that between raters was 1.3488; the antilog$S_w^{2.77}$ within raters was 1.5164, and that between raters was 1.5135. These values were used to calculate the upper and lower bounds of the expected variability for any observed value. For example, the expected true value for a person who performs the TUG in 20 seconds is (20/1.3507) = 14.8 seconds for the lower bound and (20\times1.3507) = 27 seconds for the upper bound; for a TUG performed in 30 seconds, the lower bound is 22.2 seconds and the upper bound is 40.5 seconds, and for a TUG performed in 60 seconds, the lower bound is 44.4 seconds and the upper bound is 81 seconds. Consequently, the expected ranges

---

* SPSS Inc, 233 S Wacker Dr, Chicago, IL 60606.

---
of the variability for repeated measurements are between \((20/1.5164) = 13.2\) seconds and \((20 \times 1.5164) = 30.3\) seconds for a TUG performed in 20 seconds, between 19.8 and 45.5 seconds for a TUG performed in 30 seconds, and between 39.6 and 91 seconds for a TUG performed in 60 seconds. The results of analyses of the variabilities expected for single and repeated measurements are shown in Figure 2 and Table 3. The variabilities within raters and between raters were similar (intrarater data not shown in Fig. 2).

Discussion
The present study showed that despite high ICCs, the measurement error of TUG was substantial at the individual level for older people dependent in ADL and living in residential care facilities. No connection was found between the difference in times between 2 TUG performances and the level of cognitive function when verbal cuing was permitted during testing. Instead, we found a relationship between this difference and the time needed to perform the TUG. Thus, our hypothesis that cognitive impairment would increase the variability of TUG was falsified in this group of older subjects with multiple impairments. The participants with slower TUG performance also were more likely to be cued during their TUG performance. Lower cognitive functioning was related significantly to cuing only on the third test occasion (TUG C). The measurement errors within raters and between raters were similar, and no systematic error was seen in the measurements.

Older people living in residential care facilities are often frail. They have multiple diagnoses, their intake of medications is high, they lack general strength, and they are susceptible to sickness and disease. This multisystem involvement most certainly provides sources of variability and can be attributed to a condition of unstable disability or frailty. Campbell and Buchner described frailty as "a condition or syndrome which results from a multisystem reduction in reserve capacity to the extent that a number of physiological systems are close to, or past, the threshold of symptomatic clinical failure." Frailty is considered to be the root cause of unstable disability. Movement involves an interaction between multiple systems, such as sensory input, neuromuscular function, and cognitive processes, and requires an ongoing adaptation to tasks and environment, anticipation, and planning. Despite the fact that the TUG consists of basic everyday movements, its various components can be complicated.

First, to be able to rise from a sitting position to a standing position requires both strength and technique. Walking a path for 3 m includes both acceleration and deceleration as participants prepare for a turn. The turning sequence is challenging for older people with balance disorders. Finally, turning around to sit down challenges both balance and orientation in adapting the body position to the chair. It has been suggested that a within-subject variation of 15% may be considered acceptable when repeated measurements of physical performance are obtained for young subjects who are healthy as well as middle-aged people with neck or back pain symptoms and middle-aged people who are healthy. An individual patient variability of up to 25% has been found for gait speed in patients 2 to 6 years after stroke. Our participants had an even higher degree of variability. This variability could be a consequence of the participants’ high degree of unstable disability and the combination of numerous sequential movements in the TUG.

The ICC is strongly affected by the range of scores used to calculate the coefficient; ICC is high when the difference in scores between measurements is small in comparison with the range of scores between the studied...
participants. There is no universal agreement on the interpretation of a correlation coefficient. A variety of guidelines are suggested in the literature: >.75 equals “excellent reliability,” \(^{30}\) ≥ .80 is “very reliable,” \(^{40}\) and >.75 indicates “good reliability.” \(^{41}\) Rockwood et al\(^{13}\) reported ICCs of .56 for people with cognitive impairment and .50 for people without cognitive impairment. However, several factors may have contributed to these results, as there was an average of 112±72 days between test and retest, a design not well suited for evaluating the repeatability of a measure. Thomas and Hageman\(^{14}\) reported an ICC of .87 for their small sample (n=9) of subjects with cognitive impairment. Our analyses showed high ICC(1.1) values of .91 and .92, with the lowest bound of the 95% CI being .86; there was, however, a wide range of variability in absolute figures. We wanted to provide clinicians with means to interpret the results of TUG performance at an individual level. Our results for a single measure, \(\text{antilog}S_w^{1.96}\), and for repeated measures, \(\text{antilog}S_w^{2.77}\), can be used to calculate the expected variability of any observed value, as described in the “Results” section, for a frail older person dependent in ADL and living in a residential care facility. The exact level of acceptance for measurement errors seems to be impossible to state. It depends on the purpose of the measurement; therefore, determination of the reliability of a measurement is a clinical decision. Conclusions regarding assessments cannot be made without knowledge of the expected range of variability of scores for the TUG or any other instrument.

For clinical individual-based interpretation of an intervention, the results should exceed the boundaries of variability found in the present study to be interpreted as a true change. As an example, we can apply our findings to an individual patient who performs the TUG in 50 seconds. We can expect this person’s true value to be somewhere between 37 and 67.5 seconds for a single measurement. If our intention is to use the TUG measure before treatment and after treatment for this subject, the final assessment would have to be under 33 seconds in order to be interpreted as a true positive change or above 75.8 seconds to be interpreted as a true negative change. The size of this range is so wide that it would seem impossible to expect such an improvement for a person who requires this amount of time to perform the TUG. Thus, the use of the TUG for the evaluation of treatment of an individual patient, regardless of the cognitive level, is questionable when the performance is slow. However, the TUG still could be used to provide a description of the patient’s ability, indicated by the time needed to perform the TUG.

The TUG is easily applicable even for a person with cognitive impairment. We allowed and recorded additional verbal cuing during TUG performance, as we assumed that some subjects with cognitive dysfunctions would need such cuing. Instead, we found that it was participants with slower TUG performance who were more likely to be cued at the first 2 sessions. It was not until the third test session that cognitive function, as measured with the MMSE, was related independently to the presence of verbal cuing. Therefore, despite the fact that participants were tested for the third time within 1 week, those with lower cognitive functions lacked recognition of the situation, whereas participants with higher cognitive levels, at that time, most likely gave an impression of confidence in performing the TUG.

The TUG has been recommended as a screening tool for identifying older people who are at risk for falling. The TUG times from 13 to 16 seconds have been reported as cutoff values for predicting falls in community-dwelling older people. \(^{42,43}\) With regard to our findings of variability, it seems to be difficult to state a fixed time limit as a cutoff between a person who is at risk for falling and a person who is not at risk for falling when the TUG is used at an individual level in a population of frail older people dependent in ADL and living in residential care facilities. Even with a fast TUG performance of 10 seconds, the expected true value could be within a range of 7.4 to 13.5 seconds. The causes of falls are a complex phenomenon, and the characteristics of our participants reveal that, in many ways, they are at high risk of sustaining falls. Therefore, the usefulness of a certain cutoff for the TUG could be questioned, but this notion warrants further investigation for this population.

A way in which to minimize or stabilize expected variability in physical performance assessments may be to...
use the mean of several repeated measurements. The mean of 3 repeated measurements of manually timed gait, assessed at the same test session, was reported to represent the highest level of reliability in a frail older population. However, this conclusion was based on ICC statistics. Therefore, the degree of variability remains unknown to the reader, and further studies are warranted.

The methodological framework was considered in order to create optimal conditions for the evaluation. Instructions on how to perform the TUG were thorough, and all physical therapists involved in testing procedures were trained before the assessments. We managed to conduct the retest at approximately the same time as the preceding test. Multiple pairs of therapists made the assessments, and the physical therapists were unaware of the previous results. This strategy acknowledged that the rater is a component of the test and allowed the subject-rater source of variability to influence the results. This strategy reflects the use of the tool in the clinical setting. Thus, our results reflect a combination of instrument errors, tester errors, and true subject variability. We consider these design conditions to be important for the clinical interpretation of TUG performance.

We allowed verbal cuing during TUG performance because, on the basis of our clinical experience, we assumed that a person with cognitive impairment might need such cuing to accomplish the task. Consequently, cuing could be considered to be compensatory for cognitive impairment and thus possibly necessary for the valid use of the TUG for people with cognitive deficits. However, our results revealed that cuing also was given to slow performers and thus may have been used as an encouragement to proceed in the test. The use of “subtle verbal guidance” during TUG assessments in people with dementia has been mentioned by Thomas and Hageman. In contrast, the occurrence of cuing in TUG assessments has not been reported elsewhere; therefore, we do not know whether and to what extent cuing is used in reality. In the present study, we simply recorded whether cuing was present or not; therefore, we cannot provide an exact description of the cues that were actually used. We can conclude only that verbal cuing was sometimes required for the assessment of TUG scores in this population. Whether such cuing affects the validity of TUG scores warrants further evaluation.

We believe that our findings are representative for a population of frail older people with multiple impairments, dependent in ADL, and living in residential care facilities. However, the present study did not include participants with MMSE scores of <10. Thus, our findings are not applicable to people with MMSE scores of <10.

Conclusion

The measurement error of a TUG assessment is substantial for frail older people with multiple impairments who are dependent in ADL and living in residential care facilities, regardless of the level of cognitive function, when verbal cuing is permitted during testing. The variability increases with the time needed to perform the TUG. Despite high ICCs, the ranges of expected variability can be wide and are similar within and between raters. Physical therapists should be aware of this variability before they interpret the TUG score for an individual.

References


