The reliability of six physical performance tests in older people with dementia

Christiaan G. Blankevoort¹, Marieke J.G. van Heuvelen¹, Erik J.A. Scherder¹,²

¹ Institute of Human Movement Sciences, University Medical Center Groningen, The Netherlands
² Faculty of Neuropsychology, VU University, Amsterdam, The Netherlands

Physical Therapy, 2013, 93(1):69-78
Abstract
Background: Physical performance tests are important to assess the effect of physical activity interventions in older people with dementia, but their psychometric properties have not been systematically established within this specific population.

Objective: To determine the relative and absolute test-retest reliability of the 6-meter walk test, the figure-of-eight (FoE), the Timed-up-and-Go (TUG), the FICSIT-4, the Chair-Rise test (CRT), and the Jamar dynamometer. These tests are used to assess gait speed, dynamic balance, functional mobility, static balance, lower-limb strength, and grip strength, respectively.

Design: A prospective, non-experimental study.

Methods: Older people with dementia (n=58; age range 70-92 years) performed each test at baseline and again after one week. Intraclass correlation (ICC), standard error of measurement (SEM), minimal detectable change (MDC) and (log transferred) limits of agreement of Bland-Altman plots were calculated.

Results: The relative reliabilities of the FoE, TUG, and Jamar dynamometer were excellent (ICC .90-.95) and good for the 6-meter walk test, FICSIT-4, and CRT (ICC .79-.86). SEMs and MDCs were large for all tests. The absolute reliability of the TUG and CRT was significantly influenced by the level of cognitive functioning (as assessed with the MMSE).

Limitations: The specific etiology of dementia was not obtained.

Conclusions: The results show that when used in older people with mild to moderate dementia the physical performance tests evaluated are useful to detect differences in performance between individuals and therefore suitable for cross-sectional or controlled intervention studies. They appear less suitable to monitor clinically relevant intra-individual performance changes. Future studies should focus on the development of more sensitive tests and the identification of criteria for clinically relevant changes in this rapidly growing population.
3.1 Introduction

In the next few decades, the number of people with dementia will increase dramatically (Mura, Dartigues et al. 2010). Dementia does not only lead to cognitive deficits, but also to a decline in physical performance (Kido, Tabara et al. 2010, Leandri, Cammisuli et al. 2009). Together, these declines will result in a reduction in the capacity to perform instrumental activities of daily living, such as household activities, and eventually also more basic everyday activities, such as bathing, eating and dressing (Wennie Huang, Perera et al. 2010). The ability to perform these activities are essential to a person’s autonomy and, consequently, to his or her quality of life (Iavarone, Milan et al. 2007).

Unfortunately, dementia cannot be cured, but the decline in physical performance can be slowed by physical activity interventions (Blankevoort, van Heuvelen et al. 2010). Physical performance can be considered a construct describing basic abilities that are needed to accomplish physically demanding tasks, with mobility, balance and strength as the underlying domains (Rydwik, Frandin et al. 2004). These domains can be assessed using speed measures or tasks assessing functional mobility (Podsiadlo, Richardson 1991, Kuiack, Campbell et al. 2004, Rolland, Pillard et al. 2007, Ries, Echternach et al. 2009), dynamic balance (e.g., balance during walking; Johannson, Jarnlo 1991, Shkuratova, Morris et al. 2004) and static balance (Rossiter-Fornoff, Wolf et al. 1995), and tests measuring upper limb (Resnick, Gruber-Baldini et al. 2009, Mathiowetz, Weber et al. 1984, van Heuvelen, Kempen et al. 2000) and lower limb strength (Jones, Rikli et al. 1999).

To measure the effect of exercise on these three domains in people with dementia a set of suitable and feasible tests is needed. Within the scope of the present study, this implies that the tests also need to be suitable for older people with varying degrees of cognitive impairment. Test instructions should therefore be simple, and the tests easy to administer, perform, score, and interpret, as well as being cost effective. Crucially, the tests also need to be reliable to ensure that changes in test scores reflect changes in performance and are not caused by variability in the test. Apart from fatigue and learning effects, the reliability of such tests is also assumed to be influenced by the characteristics of the individual being assessed, such as age, gender, and level of cognitive impairment (Ries, Echternach et al. 2009, Baumgartner, Jackson et al. 2003).

In the current study we evaluate the reliability of six widely used physical performance tests in older people diagnosed with dementia. Specifically, the focus of our investigations is on examining the tests with regard to their relative reliability (in terms of consistency
of within-group position; Ries, Echternach et al. 2009, Bruton, Conway et al. 2000) and absolute reliability (as reflected in the degree of variation between repeated measurements; Nordin, Rosendahl et al. 2006, van Iersel, Benraad et al. 2007). There are several reasons for this specific focus. First, there is evidence to suggest that cognitive impairment affects the reliability of different measurements (Phillips, Chu et al. 1993). Second, there are few studies that have tested the reliability of common tests in our population of interest, with two studies solely examining their relative reliabilities in small and selective samples (Thomas, Hageman 2002, van Iersel, Benraad et al. 2007). The study by Ries and colleagues (Ries, Echternach et al. 2009) is the only study that systematically evaluated the reliability of functional mobility and endurance outcomes in older people with Alzheimer's disease. The authors report large between-subject variability and recommend minimal detectable change (MDC) scores at the 90% confidence interval to monitor performance and treatment outcomes (Ries, Echternach et al. 2009).

The twofold goal of our study accordingly is to investigate the relative and absolute test-retest reliabilities of six common physical performance tests gauging mobility, balance, and strength in a group of older people with dementia, while analyzing the effect of cognitive impairment on the reliability measures, and to provide and address the relevance of MDC scores for all outcome measures.

3.2 Methods

Participants
Our study was approved by the local medical ethics committee. If participants were eligible for participation, informed consent was obtained from their legal representatives. A total of 58 participants were recruited between 2009 and 2011, from six different nursing homes and two day-care centers around the city of Groningen, the Netherlands. The study started within two months of the initial selection, during which time informed consent was obtained and assessments organized and scheduled. All participants were 70 years or older and diagnosed with dementia by the national ‘care indication center’ (CIZ), whose diagnosis and referral are mandatory in order to gain access to special geriatric care in the Netherlands. The diagnostic criteria from the CIZ are identical to the DSM-IV criteria for dementia. Exclusion criteria were a score of 9 or lower on the Mini Mental-State Exam (MMSE; Folstein, Folstein et al. 1975) to prevent measurement errors based
on the incapacity to comply with the protocol (Nordin, Rosendahl et al. 2006; Tappen, Roach et al. 1997), vision problems hampering mobility or test performance, a history of psychiatric illness (e.g., schizophrenia or bipolar disorder), neurological illness (e.g., stroke or epilepsy), alcoholism, systematic or other brain diseases that could account for the cognitive impairment, the use of a wheelchair for mobility or physical problems that could affect physical performance (e.g., a sprained ankle or (severe) musculoskeletal disorders).

**Physical performance tests**

The participants performed the assessments of gait speed, functional mobility and dynamic balance twice during each of the two test sessions, all without practice trials.

*Gait speed* was measured using the *6-meter walk test* (Thomas, Hageman 2002), which requires subjects to walk six meters in a straight line in their normal pace. The use of assistive walking devices was allowed. The outcome measure was the mean duration of two attempts, converted to walking speed (m/s), with higher scores indicating better performance. The relative reliability of the 6-meter walk test has previously been demonstrated to be excellent (ICC = .92) in older women with moderate dementia (MMSE 17.79, SD 7.17; Thomas, Hageman 2002).

*Dynamic balance* was assessed with the *Figure-of-Eight test* (FoE; Johannson, Jarnlo 1991, Pettersson, Engardt et al. 2002, Frandin, Sonn et al. 1995), which requires participants to walk two laps of a standard 10-meter long figure-of-eight course (with 15-centimeter wide contours). They are instructed to walk as fast and follow the contours as accurately as possible (Shkuratova, Morris et al. 2004). The fastest of two attempts and thus the best performance was noted (Tegner, Lysholm et al. 1986). As far as we know, the reliability of the FoE has not been investigated in older people with dementia, but two previous studies did demonstrate that in cognitively healthy older people its relative reliability was excellent (ICC = .92, and ICC = .98, respectively; Jarnlo, Nordell 2003, Helbostad, Sletvold et al. 2004).

*Functional mobility* was evaluated with the *Timed Up & Go test* (TUG; Podsiadlo, Richardson 1991) requiring participants to stand up from a chair, walk three meters, turn around, walk three meters back, and sit down again in the same chair, all in their normal pace. The use of hands and normal walking aids were allowed. The outcome measure was the mean in seconds of two trials, with faster scores indicating better performance. The TUG is reliable and valid for quantifying functional mobility (Eggermont, Gavett...
et al. 2010, Podsiadlo, Richardson 1991) and has also been found to be reliable in older people with Alzheimer's disease (ICC ≥ .95; SEM = 2.48; MDC = 4.86; Ries, Echternach et al. 2009). We included the TUG to allow comparison with the Ries et al. study (Ries, Echternach et al. 2009).

**Static balance** was gauged with the *Frailty and Injuries: Cooperative Studies of Intervention Techniques* (FICSIT-4; Rossiter-Fornoff, Wolf et al. 1995). The participants were asked to adopt four different stances, i.e. the parallel, semi-tandem, tandem, and single-leg stance, with their eyes open and without assistive devices and to try and maintain each stance for 10 seconds, with stances being sequentially adopted. The FICSIT-4 scale score ranges from 0 to 5 (0 for unsuccessful and 1 for successful parallel stance; 2 for semi-tandem stance, 3 if tandem stance was maintained less than 10 seconds, 4 for tandem stance, 5 for one legged stance). If a participant maintained the parallel or semi-tandem stance less than 10s but more than 3s an additional 0.5 points were awarded (Rossiter-Fornoff, Wolf et al. 1995). Higher scores thus indicate better performance. The FICSIT-4 showed a moderate reliability ($r = .66$; Rossiter-Fornoff, Wolf et al. 1995) in healthy older people, with pre- and post-tests scheduled three to four months apart. To our knowledge the scale has to date not been studied in older people with dementia.

**Lower limb strength** was assessed with a modified version of the 30-second sit-to-stand test from the 'Senior Fitness Test' (Rikli, Jones 2007). To prevent misinterpretation with the original test, we have labeled our edition as the 'Chair-Rise' test (CRT). We asked our participants to rise from the chair, stand up straight and sit down again as often as possible within 30 seconds (Rikli, Jones 2007, Jones, Rikli et al. 1999, Santana-Sosa, Barriopedro et al. 2008), to minimize anxiety and to prevent differences in the execution of this test, and to maximize between-subject comparisons - and thus in contrast to the original protocol - our participants were allowed to use their hands when rising. The total number of sit-to-stands (Rikli, Jones 2007) constituted the outcome score, with higher scores indicating better performance. The original sit-to-stand test (Rikli, Jones 2007) showed a good relative reliability amongst cognitively healthy older people (ICC = .84, ICC = .92, for male and female subjects, respectively; Jones, Rikli et al. 1999) and has, to our knowledge, not been studied in older people with dementia.

**Grip strength** was measured with a *Jamar dynamometer.* While standing and holding the dynamometer in their dominant hand, with the arm extended and the palm of their
hand facing their leg, the participants were instructed to squeeze the grip as hard as possible. The strongest of three attempts (in kilograms) was recorded, with higher values reflecting better performance. The relative reliability of grip strength as measured with the Jamar dynamometer was earlier found to be excellent in cognitively unimpaired elders (ICC = .92; Bohannon, Schaubert 2005) but moderate (ICC = .72) in older people with dementia (Thomas, Hageman 2002).

**Global cognitive functioning**

The participants’ global cognitive abilities were assessed by the primary researcher (C.G.B.), who is a trained neuropsychologist, using the Mini-Mental State Examination (MMSE; Folstein, Folstein et al. 1975). All patients were assessed in the week prior to their first physical test. Scores on the MMSE range from 0-30, with a score below 10 being indicative of severe cognitive impairment, while scores between 10-19 and those between 20-24 reflect moderate and mild cognitive impairment, respectively (Binetti, Mega et al. 1998, Kapaki, Paraskevas 2005).

**Procedure**

For the practical approaches to optimize the communication with our participants, we refer to the extensive description Ries and colleagues provided in their study of Alzheimer patients (Ries, Echternach et al. 2009). In short, creating a relaxed, pleasant atmosphere and using simple commands were key elements. Each assessment was first demonstrated to the patient and, if necessary, cues or gestures were provided (Ries, Echternach et al. 2009). To keep test conditions comparable, variations in staff training, time of day, location, and sequence of tests were kept to a minimum. To prevent bias examiners were blinded for previous test scores and, if possible, for the level of cognitive functioning.

All participants performed the six physical tests in the same sequence at baseline and at the second session scheduled one week later. The tests were all administered in the patients’ own nursing homes or day-care centers by five trained bachelor and master students from the Human Movement Sciences program of the Center of Human Movement Sciences, University Medical Center Groningen, The Netherlands.

Two of the test sites had insufficient space for the FoE, 12 participants did not perform this test. Another six participants were unable to perform the CRT due to arthritis, knee operations, or other knee problems. One participant could not perform the grip-strength test because of failure of the equipment.
Data Analyses

The data were analyzed using SPSS 16.0 for Windows† and Excel 2003 for Windows‡. First, the data were analyzed for skewness, kurtosis, and heteroscedasticity using the Koenker test. When necessary ($p < .05$), the data were log transformed. Relative test-retest reliability was calculated with the Intraclass Correlation Coefficient (ICC), which reflects the consistency to which the within-group position is maintained (Ries, Echternach et al. 2009, Bruton, Conway et al. 2000). The ICC was calculated using the two-way random, absolute agreement on single measures model with a 95% confidence interval. An ICC above 0.70 is deemed sufficient for group comparison, but for individual monitoring the ICC should exceed .90-.95 (Scientific Advisory Committee of the Medical Outcomes Trust 2002).

Even with a high ICC, the trial-to-trial consistency of physical measurements can be poor, especially in heterogeneous datasets (Bruton, Conway et al. 2000, Nordin, Rosendahl et al. 2006, van Iersel, Benraad et al. 2007). We thus also considered their absolute reliabilities (Nordin, Rosendahl et al. 2006, van Iersel, Benraad et al. 2007), which we calculated with the Bland-Altman 95% limits of agreement and the standard error of measurement (SEM; Bruton, Conway et al. 2000, Bland, Altman 1986, Keating, Matyas 1998). To facilitate interpretation of the results, the SEM is reported in the same quantity used for the original measurement (e.g., in kilograms, m/s (speed), time(s)). It thus provides the range within which a participant's true score may fall (Domholdt 2005). If the SEM is small, indicating high absolute reliability, the true score is close to the recorded score (Bruton, Conway et al. 2000). The probabilities of the normal curve can then be applied to the SEM (Ries, Echternach et al. 2009), which means that with a probability of 68% the score on a next assessment will be within 1 SEM from the original score. Moreover, with a probability of 95% the next score for the same participant will be within 2 SEMs from the first score. The following formula was used (Bruton, Conway et al. 2000):

$$SEM = sd * \sqrt{(1 - ICC)}$$

† SPSS Inc. 233 S Wacker Dr, Chicago Il 60606-6412, USA
‡ Microsoft cooperation, One Microsoft Way, Redmond, WA 98052-7329, USA
The 95% confidence intervals for the SEM were calculated as described by Stratford and Goldsmith (Stratford 1997):

$$\left[ \frac{SSE}{\chi^2 \alpha, df_e} - \frac{SSE}{\chi^2 1 - \alpha, df_e} \right]$$

The abbreviations in the latter formula have the following meaning: SSE, the sum of squared errors in the ANOVA table, $\chi^2_{\alpha, df_e}$ the chi-square value for probability level $\alpha$, and $df_e$ the degrees of freedom of the SSE provided in the ANOVA table (Stratford 1997). The square roots of these two values provide the borders for the 95% CI of the SEM (Stratford 1997).

Finally, to be able to interpret changes in test scores, the minimal detectable change (MDC) with 95% confidence interval (CI) was calculated (Ries, Echternach et al. 2009):

$$MDC_{95} = SEM \times Z_{95}(1.96) \times \sqrt{2}$$

The MDC is the required magnitude of observable change that exceeds the anticipated measurement error and within-subject variability (Haley, Fragala-Pinkham 2006). In other words, if a participant’s score exceeds the value of the MDC than it can be said to reflect a true change in performance with 95% confidence.

The calculations were performed for the total group and stratified by level of cognition, distinguishing between participants with mild (MMSE ≥ 20) and those with moderate (MMSE 10-19) cognitive impairment (Binetti, Mega et al. 1998, Kapaki, Paraskevas 2005). No overlap in the CI of the ICC or the SEM was taken to indicate a statistically significant difference in performance scores for the groups with mild and moderate decline (Flechner, Tseng 2011).

For a visual inspection of the similarity between the two measurements Bland-Altman plots were created with the limits of agreements (LoA). For non-skewed data the following formula was used to calculate the LoA (Euser, Dekker et al. 2008):

$$\text{Mean difference} \pm 1.96 \times SD$$

For skewed data the following formula was used to calculate the LoA (Euser, Dekker et al. 2008):

$$\frac{-2 \bar{X}(10^a - 1)}{(10^a + 1)} + \frac{2 \bar{X}(10^a - 1)}{(10^a + 1)} ; \alpha = 1.96 \times \sqrt{2\sigma_{Err}^2}$$

With $\sigma_{Err}^2$ reflecting the residual-error variance.
Table 3.1. Characteristics of the Participants.

<table>
<thead>
<tr>
<th></th>
<th>Total Group (n=58)</th>
<th>Mild Cognitive Impairment (n=30)</th>
<th>Moderate Cognitive Impairment (n=28)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (SD)</td>
<td>82.47 (5.31)</td>
<td>82.37 (5.16)</td>
<td>82.57 (5.55)</td>
<td>.89a</td>
</tr>
<tr>
<td>Min-Max</td>
<td>70-91</td>
<td>70-91</td>
<td>70-92</td>
<td></td>
</tr>
<tr>
<td>Female (♀)</td>
<td>41 (70.7%)</td>
<td>21 (70.0%)</td>
<td>21 (75%)</td>
<td>.49b</td>
</tr>
<tr>
<td>Place of residence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH</td>
<td>34%</td>
<td>NH 34%</td>
<td>NH 46%</td>
<td>.01b</td>
</tr>
<tr>
<td>HE</td>
<td>12%</td>
<td>HE 3%</td>
<td>HE 21%</td>
<td></td>
</tr>
<tr>
<td>HL</td>
<td>53%</td>
<td>HL 73%</td>
<td>HL 32%</td>
<td></td>
</tr>
<tr>
<td>MMSE M (SD)</td>
<td>19.24 (4.37)</td>
<td>22.77 (2.13)</td>
<td>15.46 (2.63)</td>
<td>&lt;.001a</td>
</tr>
<tr>
<td>Min-Max</td>
<td>10-28</td>
<td>20-28</td>
<td>10-19</td>
<td></td>
</tr>
<tr>
<td>Walking Aid (Yes)</td>
<td>26 (44.8%)</td>
<td>15 (50.0%)</td>
<td>11 (39.3%)</td>
<td>.41b</td>
</tr>
</tbody>
</table>

*a*-test
*b* Chi-square test

NH, Nursing Home; HE, Home for the Elderly; HL, Home Living; MMSE, Mini Mental-State Examination; Min, Minimum; Max, maximum.

3.3 Results

Table 3.1 presents the characteristics of the 58 participants in the final sample; 17 participants were male and 41 female, with ages ranging from 70 to 92 years. No significant differences in age, gender, or the use of walking aids were found between the participants with mild (MMSE: 20-28) and moderate (10-19) cognitive impairment. However, the differences for place of residence were statistically significant.

Table 3.2 presents the relative and absolute reliability values for the six physical performance tests for the total group. The relative reliabilities of the FoE, the TUG, and Jamar dynamometer were excellent (ICC > .90), and those for the 6-meter walk test, the CRT, and the FICSIT-4 good (ICC = .75-90). The width of the confidence intervals (CI) of the ICCs ranged between .05 and .20, with the TUG having the smallest CI and the FICSIT-4 the largest. The absolute reliabilities of the tests, measured with the SEMs and MDCs, were large.

Figure 3.1 shows the Bland-Altman plots with the 95% limits of agreement for the six tests calculated for the total group (Bland, Altman 1986, Euser, Dekker et al. 2008). The data of the FoE, the TUG, and Jamar dynamometer were positively skewed and heteroscedastic, with higher means yielding higher variability, as is reflected by the wider LoAs. The data of the 6-meter walk test, the CRT, and the FICSIT-4 were homoscedastic.
and, consequently, had a constant LoA.

Table 3.3 lists the test scores and reliability values as a function of cognitive functioning (assessed with the MMSE). The CRT was the only test yielding a significant group difference, with participants with a milder cognitive deficit achieving better scores. We found no significant between-group difference for relative reliability, but the absolute reliabilities of the TUG and CRT did show a significant difference, as reflected in their elevated MDCs. The MDC of the TUG was smaller (3.96 sec) in persons with mild cognitive impairment versus those with moderate cognitive impairment (8.07 sec). The MDC of the CRT was larger (4.21 stands) in persons with mild cognitive impairment versus those with moderate cognitive impairment (2.30 stands).
Table 3.2. Descriptive and Reliability Measures of the Physical Performance Tests in the Study Group Based on a One-Week Test-Retest Interval.

<table>
<thead>
<tr>
<th>Test</th>
<th>n</th>
<th>Mean (SD) Test</th>
<th>Mean (SD) Retest</th>
<th>KT</th>
<th>F-value</th>
<th>p</th>
<th>ICC</th>
<th>CI95 ICC</th>
<th>SEM</th>
<th>CI95 SEM</th>
<th>MDC95</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-meter walk test (m/s)</td>
<td>58</td>
<td>.77 (.25)</td>
<td>.75 (.28)</td>
<td>.42</td>
<td>-.48</td>
<td>.49</td>
<td>.86</td>
<td>.78-.92</td>
<td>.10</td>
<td>.08-.12</td>
<td>.27</td>
</tr>
<tr>
<td>Minimum-Maximum</td>
<td></td>
<td>.32-.1.60</td>
<td>.24-1.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FoE (s)</td>
<td>46</td>
<td>45.97 (21.23)</td>
<td>45.51 (20.90)</td>
<td>.00</td>
<td>.14*</td>
<td>.71</td>
<td>.91</td>
<td>.85-.95</td>
<td>6.26</td>
<td>5.41-8.21</td>
<td>17.35</td>
</tr>
<tr>
<td>Minimum-Maximum</td>
<td></td>
<td>19.26-120.00</td>
<td>17.30-114.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUG (s)</td>
<td>58</td>
<td>18.55 (9.74)</td>
<td>18.68 (9.01)</td>
<td>.00</td>
<td>.99*</td>
<td>.32</td>
<td>.94</td>
<td>.92-.97</td>
<td>2.12</td>
<td>1.74-2.52</td>
<td>5.88</td>
</tr>
<tr>
<td>FICSIT-4 (pts)</td>
<td>58</td>
<td>2.55 (1.10)</td>
<td>2.58 (1.32)</td>
<td>.84</td>
<td>.06</td>
<td>.80</td>
<td>.79</td>
<td>.67-.87</td>
<td>.55</td>
<td>.47-.69</td>
<td>1.52</td>
</tr>
<tr>
<td>Minimum-Maximum</td>
<td></td>
<td>0.00-4.00</td>
<td>0.00-5.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chair-Rise test (#)</td>
<td>52</td>
<td>8.12 (2.95)</td>
<td>8.30 (3.32)</td>
<td>.87</td>
<td>.54</td>
<td>.47</td>
<td>.84</td>
<td>.73-.90</td>
<td>1.26</td>
<td>1.06-1.57</td>
<td>3.49</td>
</tr>
<tr>
<td>Minimum-Maximum</td>
<td></td>
<td>2.00-14.50</td>
<td>2.00-18.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jamar dynamometer (kg)</td>
<td>57</td>
<td>20.77 (9.18)</td>
<td>20.55 (8.34)</td>
<td>.01</td>
<td>.01*</td>
<td>.95</td>
<td>.90</td>
<td>.84-.94</td>
<td>2.74</td>
<td>2.05-2.98</td>
<td>7.59</td>
</tr>
<tr>
<td>Minimum-Maximum</td>
<td></td>
<td>9.00-55.00</td>
<td>10.00-46.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*ICC calculation over log-transferred data.
KT, Koenker test for heteroscedasticity; df, degrees of freedom; ICC, Intra-class correlation; CI95, 95% Confidence Interval; SEM, Standard error of measurement; MDC, Minimal detectable change; FoE, Figure of Eight; TUG, Timed Up-and-Go test, FICSIT-4, Frailty and Injuries: Cooperative Studies of Intervention Techniques-4; pts, points.
Table 3.3. Baseline and Retest Outcomes, and Reliability Values for the Six Physical Performance Tests Stratified By Current Cognitive Functioning.

<table>
<thead>
<tr>
<th></th>
<th>6-meter walk test (m/s)</th>
<th>FoE (s)</th>
<th>TUG (s)</th>
<th>FICSIT-4</th>
<th>Chair-Rise test (#)</th>
<th>Jamar dynamometer (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mild cognitive impairment (MMSE 20-28)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>30</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Mean session 1 (SD)</td>
<td>.74 (.26)</td>
<td>48.12 (25.21)</td>
<td>16.95 (7.49)</td>
<td>2.72 (1.14)</td>
<td>9.12 (3.11)*</td>
<td>20.83 (7.87)</td>
</tr>
<tr>
<td>Mean session 2 (SD)</td>
<td>.73 (.30)</td>
<td>45.61 (24.93)</td>
<td>17.01 (6.96)</td>
<td>2.83 (1.29)</td>
<td>9.33 (3.56)**</td>
<td>20.97 (6.84)</td>
</tr>
<tr>
<td>ICC (CI&lt;sub&gt;95&lt;/sub&gt;)</td>
<td>.83 (.67-.91)</td>
<td>.94 (.86-.97)</td>
<td>.96 (.92-.98)</td>
<td>.82 (.65-.91)</td>
<td>.79 (.60-.90)</td>
<td>.86 (.72-.93)</td>
</tr>
<tr>
<td>SEM (CI&lt;sub&gt;95&lt;/sub&gt;)</td>
<td>.11 (.09-.11)</td>
<td>6.24 (5.63-10.03)</td>
<td>1.43 (1.06-1.79)</td>
<td>.59 (.48-.81)</td>
<td>1.52 (1.22-2.08)</td>
<td>2.75 (1.85-3.15)</td>
</tr>
<tr>
<td>MDC&lt;sub&gt;95&lt;/sub&gt;</td>
<td>.29</td>
<td>17.30</td>
<td>3.96</td>
<td>1.64</td>
<td>4.21</td>
<td>7.62</td>
</tr>
<tr>
<td><strong>Moderate cognitive impairment (MMSE 10-19)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>28</td>
<td>21</td>
<td>28</td>
<td>28</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>Mean session 1 (SD)</td>
<td>.80 (.25)</td>
<td>43.42 (15.41)</td>
<td>20.26 (11.59)</td>
<td>2.38 (1.04)</td>
<td>6.85 (2.21)*</td>
<td>20.71 (10.52)</td>
</tr>
<tr>
<td>Mean session 2 (SD)</td>
<td>.78 (.26)</td>
<td>45.51 (15.39)</td>
<td>20.46 (10.63)</td>
<td>2.30 (1.31)</td>
<td>7.00 (2.49)**</td>
<td>20.13 (9.77)</td>
</tr>
<tr>
<td>ICC (CI&lt;sub&gt;95&lt;/sub&gt;)</td>
<td>.89 (.78-.95)</td>
<td>.85 (.67-.94)</td>
<td>.94 (.87-.97)</td>
<td>.80 (.61-.90)</td>
<td>.88 (.73-.95)</td>
<td>.94 (.87-.97)</td>
</tr>
<tr>
<td>SEM (CI&lt;sub&gt;95&lt;/sub&gt;)</td>
<td>.09 (.07-.13)</td>
<td>6.00 (4.01-7.58)</td>
<td>2.91 (2.10-3.61)</td>
<td>.60 (.48-.82)</td>
<td>.83 (.65-1.04)</td>
<td>2.57 (2.02-3.47)</td>
</tr>
<tr>
<td>MDC&lt;sub&gt;95&lt;/sub&gt;</td>
<td>.25</td>
<td>16.63</td>
<td>8.07</td>
<td>1.66</td>
<td>2.30</td>
<td>7.11</td>
</tr>
</tbody>
</table>

*significant difference at baseline between participants with higher and lower scores on the MMSE (p < .01); **significant difference at retest between participants with higher and lower scores on the MMSE (p < .01);
FoE, Figure of Eight; TUG, Timed Up-and-Go test, FICSIT-4, Frailty and Injuries: Cooperative Studies of Intervention Techniques-4; CI<sub>95</sub>, 95% Confidence-Interval.
Figure 3.1. Bland-Altman Plots Showing the Levels of Agreement for the Heteroscedastic and the Homoscedastic Data for the Six Tests Evaluated. The Two Measurements Were One Week Apart.
3.4 Discussion

The main goal of our study was to evaluate the relative and absolute reliability of six physical functioning tests in older people (70-92 yrs) with dementia, with a focus on tests gauging gait speed, dynamic balance, functional mobility, static balance, lower-limb strength, and grip strength. Additionally we analyzed the effects of cognitive impairment on the reliability coefficients.

Relative reliability

The results showed that the relative reliability was excellent for the TUG, FoE, and Jamar dynamometer (ICC > .90), and good for the 6-meter walk test, CRT and FICSIT-4 (ICC .75-.90). The differences in relative reliability between the participants with mild and those with moderate cognitive impairment were non-significant.

The values we obtained for the FoE, Jamar dynamometer, 6-meter walk test, and the CRT were similar to those earlier reported for similarly aged participants with (Thomas, Hageman 2002) and without dementia (Jones, Rikli et al. 1999, Jarnlo, Nordell 2003, Bohannon, Schaubert 2005). The values we recorded for the TUG were somewhat lower than those Ries et al. reported for Alzheimer patients (ICC = .985-.988; Ries, Echternach et al. 2009). It is likely that this disparity is caused by differences in the characteristics of the two patient groups. The percentage of women in our sample was higher than in the Ries study (Ries, Echternach et al. 2009).

A study solely evaluating female patients with different subtypes of dementia found lower relative reliability scores for the TUG (ICC = .87) and the dynamometer test (ICC = .70; Thomas, Hageman 2002). In general, men are stronger and have more endurance than women, and by excluding male participants the group becomes more homogeneous, decreasing the relative reliability of these tests. Accordingly, when male and female participants are considered as one single group this causes an upward bias in the reliability coefficient.

The TUG, FoE, and Jamar dynamometer values exceeded the threshold for minimal acceptable reliability (ICC = .90) and thus may be useful for individual monitoring (Ries, Echternach et al. 2009, Scientific Advisory Committee of the Medical Outcomes Trust 2002). However, for that goal their absolute reliabilities should also be considered to establish the within-subject test-retest variability, which we do in the next section.

Given their lower ICC scores, the 6-meter walk test, the CRT, and the FICSIT-4 do
not appear suitable for individual performance monitoring. However, since all six tests exceeded the threshold for group comparisons (ICC > .70; Scientific Advisory Committee of the Medical Outcomes Trust 2002), they do seem suitable for use in cross-sectional or controlled intervention studies.

**Absolute reliability**

The absolute reliability of a test provides an estimate of the precision of its outcome scores on repeated testing (Overend, Anderson et al. 2010). The standard error of measurement (SEM) and the MDC are easy to interpret because they are expressed in the same units as the original measure and are, as such, very useful for clinicians to determine individual improvement (Domholdt 2005). They conveniently allow the 95% confidence interval (2 SEMs) to be computed for the true score and the range in which a next score, from a stable participant, would be expected. The MDC is based on the SEM, but is more conservative (~2.7 SEMs). If a score change is larger than the MDC, this difference is not caused by a measurement error or patient variability (with a probability of 95%; Ries, Echternach et al. 2009). Because, the MDC and SEM are so closely linked, in this discussion we will focus on the MDC only.

To interpret the MDC correctly, the variance of the data should remain constant with increasing means (homoscedastic distribution). This was true for the 6-meter walk test, the FICSIT-4, and the CRT. It required an improvement of .27 m/s and an increase of 1.52 pts for the MDCs of the 6-meter walk test and FICSIT-4 to be exceeded. The absolute reliability of the CRT was influenced by the participants’ level of cognitive impairment. Consequently, it took an improvement of 4.21 stands (mild cognitive impairment) or 2.30 stands (moderate cognitive impairment) to exceed the MDC. It is possible that the higher absolute reliabilities for the participants with moderate cognitive impairment are explained by a floor effect.

For the FoE, the TUG, and Jamar dynamometer the variance did not remain constant with incremental means (heteroscedastic distribution; see the Bland-Altman plots). Here, the MDCs should be interpreted more cautiously. Given the heteroscedastic properties of the data, the MDC increases with an increase of the mean (as is reflected by the V-shaped lines in the Bland-Altman plots; Euser, Dekker et al. 2008). This indicates that the participants who attained lower scores on these three tests showed less variability than their peers achieving higher scores. Consequently, for the FoE, TUG, and Jamar dynamometer clinically relevant changes might not be detected as such (for low scores),
or the importance of changes might be overestimated (for high scores). This should be kept in mind when interpreting their respective MDCs.

For the FoE to exceed the MDC an improvement of 17.35 (s) was required, while improvement on the dynamometer test needed to be in excess of 7.59 (kg). The results of the TUG were affected by the participants’ cognitive abilities, requiring an improvement of 3.96 (s) for participants with mild and 8.07 (s) for those with moderate cognitive impairment. The distinction on the TUG between participants with mild and moderate cognitive impairment are in line with a study amongst Alzheimer patients.\(^{11}\)

Although the MDC should facilitate the appraisal of individual improvement on certain tests, the large margins of improvement the tests seem to require (e.g., 7.59 kg for grip strength) warrant discussion of their practical relevance. The first issue we will address is whether it is realistic to expect increases in performance larger than the MDC. And secondly, whether performance improvements lower than the MDC have any clinical relevance (which, ideally, should not be the case).

To address the first issue, the systematic review of Blankevoort and colleagues (Blankevoort, van Heuvelen et al. 2010), shows that only one study (out of sixteen) showed a post-intervention improvement larger than the MDCs for the TUG, the Sit-to-Stand test and gait and balance abilities measured with the Tinetti scale (Santana-Sosa, Barriopedro et al. 2008). This suggests that improvements exceeding the MDC are not viable and thus these tests are probably unsuitable to quantify treatment effects within this specific population.

Only a limited amount of information about clinical relevance is available. In a study of frail older adults, among whom were dementia patients, van Iersel et al. concluded that an increased walking speed of .21 m/s reduced the (expert-rated) risk of falling (van Iersel, Munneke et al. 2008). This is below the MDC computed in our study (.27 m/s), rendering gait speed, as measured with the 6-meter walk test, a less suitable measure to detect changes of this magnitude in fall risk. The more sophisticated GAITRite\textsuperscript{®} walkway system yielded a smaller MDC (.11 m/s; Ries, Echternach et al. 2009) and might be more suitable to assess clinically relevant changes in gait speed. Van Iersel and colleagues also judged an improvement of 10.1 seconds on the TUG as clinically relevant (van Iersel, Munneke et al. 2008). As this value is larger than the MDCs computed in our study, the TUG appears suitable to detect clinically relevant improvements of this magnitude (as judged by experts). Unfortunately, we were unable to compare our MDC findings on the other tests with the literature as we did not find similar studies reporting clinically relevant

In summary, we conclude that the MDCs obtained for the six physical performance tests evaluated limit their applicability to detect individual improvements in older people with mild to moderate cognitive deficits in the targeted domains as (1) the increases in performance need to be very large to exceed the MDC and (2) the MDCs may be too large to allow small, but clinically relevant changes to be detected. Future research should focus on the development of more sensitive tests to monitor physical performance and identify criteria for clinical relevant changes in this population.

Limitations
This study has several limitations. First, we were unable to retrieve the etiologies (e.g., Alzheimer’s disease or vascular dementia) of the dementia syndromes from the patients’ medical records, as diagnoses were mostly reported as ‘dementia’, or ‘dementia syndrome’. Six participants had MMSE scores higher than 24 (cut-off for mild cognitive deficit). All six participants were attending geriatric adult day care. This means that they had a diagnoses of dementia according to the DSM-IV criteria, which is necessary for approval of the national Care Indication Center for participation in geriatric adult day care. More importantly, the MMSE is a global cognitive screening instrument and thus suitable to differentiate groups, but not appropriate to diagnose individuals.

Second, we modified elements of some of the original test protocols. For example, instructions were repeated if necessary, and hand use was allowed in the CRT, our equivalent of the sit-to-stand test. These adjustments may have influenced the comparative validities of the tests. Given the correlation between upper and lower extremity strength (in our study \( r = .50 \)), it is not likely that the use of hands had a large effect on the outcome of our chair-rise test, although further research is necessary to determine the exact impact.

Third, our sample size was based on convenience and a post-hoc analysis showed that for most tests a sample of 50 subjects was required, but as 58 participants completed the test, this did not pose a problem.

Fourth, because the participants were tested at their place of residence and because examiners had to interact with the participants they could not completely be blinded from the level of cognitive functioning. The examiners did, however, not have any information regarding the MMSE scores of the participants at the moment of testing.
Finally, while the generalizability of our study appears adequate given the heterogeneity of the participants, its generalizability might be hampered by the limited geographical variability.

3.5 Conclusion

The relative reliabilities of six physical performance tests, i.e., the 6-meter walk test, FoE, TUG, FICSIT-4, CRT, and the Jamar dynamometer, were good to excellent. The tests are thus all applicable for cross-sectional and controlled intervention studies of older people with mild to moderate dementia. However, their minimal detectable change (MDC) values were large, which seriously complicates the detection of clinically relevant changes in this population. Future research should focus on the development of more sensitive tests to assess and monitor physical performance in dementia patients and to define criteria for clinically relevant changes.