Reliability of Six Physical Performance Tests in Older People With Dementia

Christiaan G. Blankevoort, Marieke J.G. van Heuvelen, Erik J.A. Scherder

Background. Physical performance tests are important for assessing 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. The purpose of this study was to determine the relative and absolute test-retest reliability of the 6-m walk test, the Figure-of-Eight Walk Test (F8W), the Timed "Up & Go" Test (TUG), the Frailty and Injuries: Cooperative Studies of Intervention Techniques–4 (FICSIT–4) Balance Test, 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. This investigation was a prospective, nonexperimental study.

Methods. Older people with dementia (n=58, age range=70–92 years) performed each test at baseline and again after 1 week. Intraclass correlation coefficients (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 reliability of the F8W, TUG, and Jamar dynamometer was excellent (ICC=.90–.95) and good for the 6-m walk test, FICSIT–4, and CRT (ICC=.79–.86). The 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 Mini-Mental State Examination [MMSE]).

Limitations. The specific etiology of dementia was not obtained.

Conclusions. The physical performance tests evaluated are useful for detecting differences in performance between older people with mild to moderate dementia and, therefore, are 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.

C.G. Blankevoort, MSc, Center for Human Movement Sciences, University Medical Center Groningen, PO Box 196, 9700 AD Groningen, the Netherlands. Address all correspondence to Mr Blankevoort at: c.g.blankevoort@med.umcg.nl.

M.J.G. van Heuvelen, PhD, Center for Human Movement Sciences, University Medical Center Groningen.

E.J.A. Scherder, PhD, Center for Human Movement Sciences, University Medical Center Groningen, and Faculty of Neuropsychology, VU University, Amsterdam, the Netherlands.

[Blankevoort CG, van Heuvelen MJG, Scherder EJA. Reliability of six physical performance tests in older people with dementia. *Phys Ther.* 2013;93:69–78.]

© 2013 American Physical Therapy Association

Published Ahead of Print: September 13, 2012
Accepted: September 4, 2012
Submitted: May 20, 2011

Post a Rapid Response to this article at:
ptjournal.apta.org

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

In the next few decades, the number of people with dementia will increase dramatically.[1] Dementia does not only lead to cognitive deficits, but also to a decline in physical performance.[2,3] Together, these declines will reduce the person's capacity to perform instrumental activities of daily living (eg, household activities) and eventually, the basic everyday activities (eg, bathing, eating, dressing).[4] The ability to perform these activities is essential to a person's autonomy and, consequently, to his or her quality of life.[5]

Unfortunately, dementia cannot be cured, but the decline in physical performance can be slowed by physical activity interventions.[6] Physical performance can be considered a construct that describes the basic abilities necessary to accomplish physically demanding tasks, with mobility, balance, and strength as the underlying domains.[7] These domains can be evaluated by using speed measures or tasks that assess functional mobility,[8–11] dynamic balance (eg, balance during walking),[12,13] and static balance[14] and tests that measure upper-limb[15–17] and lower-limb[18] strength.

In order to measure the effect of exercise on these 3 domains in people with dementia, a set of suitable and feasible tests is needed. Within the scope of the present study, "suitable and feasible" implies that the tests also need to be suitable for older people with varying degrees of cognitive impairment. Therefore, test instructions should be simple, and the tests easy to administer, perform, score, and interpret, as well as 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 assumed to be also influenced by the characteristics of the individual being assessed, such as age, sex, and level of cognitive impairment.[11,19]

In the current study, we evaluated the reliability of 6 widely used physical performance tests in older people diagnosed with dementia. Specifically, the focus of our investigations was on examining the tests with regard to their relative reliability (in terms of consistency of within-group position)[11,20] and absolute reliability (as reflected in the degree of variation between repeated measurements).[21,22] There were several reasons for this specific focus. First, there is evidence to suggest that cognitive impairment affects the reliability of different measurements.[23] Second, there are few studies that have tested the reliability of common tests in our population of interest, with 2 studies solely examining their relative reliability in small and selective samples.[22,24] The study by Ries and colleagues[11] is the only study that systematically evaluated the reliability of functional mobility and endurance outcomes in older people with Alzheimer disease. The authors reported large between-subject variability and recommended minimal detectable change (MDC) scores at the 90% confidence interval (CI) to monitor performance and treatment outcomes.[11]

The 2-fold goal of our study accordingly was to investigate the relative and absolute test-retest reliability of 6 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.

Method

Participants

Our study was approved by the local medical ethics committee. If individuals were eligible for participation, informed consent was obtained from their legal representatives. A total of 58 participants were recruited between 2009 and 2011 from 6 different nursing homes and 2 day care centers around the city of Groningen, the Netherlands. The study started within 2 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 American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition, (DSM-IV) criteria for dementia.[25] Exclusion criteria were a score of 9 or lower on the Mini-Mental State Examination (MMSE)[26] to prevent measurement errors based on the incapacity to adhere to the protocol,[21,27] vision problems hampering mobility or test performance, a history of psychiatric illness (eg, schizophrenia or bipolar disorder), neurological illness (eg, stroke or epilepsy), alcoholism, systemic or other brain diseases that could account for the cognitive impairment, or the use of a wheelchair for mobility or physical problems that could potentially affect physical performance (eg, 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 2 test sessions, all without practice trials.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

Gait speed was measured using the 6-m walk test,[24] which requires participants to walk 6 m in a straight line at their normal pace. The use of assistive walking devices was allowed. The outcome measure was the mean duration of 2 attempts, converted to walking speed (m/s), with higher scores indicating better performance. The relative reliability of the 6-m walk test has previously been demonstrated to be excellent (intraclass correlation coefficient [ICC]= .92) in older women with moderate dementia (MMSE=17.79, SD= 7.17).[24]

Dynamic balance was assessed with the Figure-of-Eight Walk Test (F8W),[12,28,29] which requires participants to walk 2 laps of a standard, 10-m-long course shaped like a figure eight (with 15-cm-wide contours). They are instructed to walk and follow the contours as fast and accurately as possible.[13] The fastest of 2 attempts, and thus the best performance, was noted.[30] To our knowledge, the reliability of the F8W has not been investigated in older people with dementia, but 2 previous studies did demonstrate that in older people who were cognitively healthy, its relative reliability was excellent (ICC=.92, and ICC=.98, respectively).[31,32]

Functional mobility was evaluated with the Timed “Up & Go” Test (TUG),[10] requiring participants to stand up from a chair, walk 3 m, turn around, walk 3 m back, and sit down again in the same chair, all at their normal pace. The use of hands and normal walking aids was allowed. The outcome measure was the mean (in seconds) of 2 trials, with faster scores indicating better performance. The TUG is reliable and valid for quantifying functional mobility[10,33] and has been found to be reliable in older people with Alzheimer disease (ICC≥.95; standard error of measurement [SEM]=2.48; minimal detectable change [MDC]=4.86).[11] We included the TUG to allow comparison with the study by Ries et al.[11]

Static balance was gauged with the Frailty and Injuries: Cooperative Studies of Intervention Techniques–4 (FICSIT–4).[14] The participants were asked to adopt 4 different stances (ie, parallel, semi-tandem, tandem, and single-leg stances) with their eyes open and without assistive devices and to try to 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 parallel stance was maintained less than 10 seconds, 4 for parallel stance, and 5 for single-legged stance). If a participant maintained the parallel or semi-tandem stance less than 10 seconds but more than 3 seconds, an additional 0.5 point was awarded.[14] Higher scores thus indicate better performance. The FICSIT–4 showed moderate reliability (r=.66)[14] in older people who were healthy, with pretests and posttests scheduled 3 to 4 months apart. To our knowledge, the scale has not been studied in older people with dementia to date.

Lower-limb strength was assessed with a modified version of the 30-second sit-to-stand test from the Senior Fitness Test.[34] To prevent misinterpretation with the original test, we 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.[18,34,35] To minimize anxiety, prevent differences in the execution of this test, and maximize between-subject comparisons, our participants (in contrast to the original protocol) were allowed to use their hands when rising. The total number of sit-to-stands[34] constituted the outcome score, with higher scores indicating better performance. The original sit-to-stand test[34] showed good relative reliability among older people who were cognitively healthy (ICC=.84 and ICC=.92, for male and female participants, respectively)[18] and has, to our knowledge, not been studied in older people with dementia.

Grip strength was measured with a Jamar dynamometer (Sammons Preston Rolyan, Bolingbrook, Illinois). 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 3 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 (ICC=.92)[36] in elderly people without cognitive impairment, but moderate (ICC=.72) in older people with dementia.[24]

Global Cognitive Functioning

The participants’ global cognitive abilities were assessed by the primary researcher (C.G.B.), who is a trained neuropsychologist, using the MMSE.[26] All participants were assessed in the week prior to their first physical test. Scores on the MMSE range from 0 to 30, with a score below 10 being indicative of severe cognitive impairment and scores between 10 and 19 and between 20 and 24 reflecting moderate and mild cognitive impairment, respectively.[37,38]

Procedure

For the practical approaches to optimize the communication with our participants, we refer to the extensive description Ries and colleagues[11] provided in their 2009 study of patients with Alzheimer dis-

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

ease. 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.[11] 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 from previous test scores and, if possible, for the level of cognitive functioning.

All participants performed the 6 physical tests in the same sequence at baseline and at the second session scheduled 1 week later. The tests were all administered at the patients' own nursing homes or day care centers by 5 trained bachelor degree and master degree 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 F8W, and 12 participants did not perform this test. Another 6 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 Analysis

The data were analyzed using SPSS 16.0 for Windows (SPSS Inc, Chicago, Illinois) and Excel 2003 for Windows (Microsoft Corporation, Redmond, Washington). 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 ICC, which reflects the consistency to which the within-group position is maintained.[11,20] The ICC was calculated using the 2-way, random, absolute agreement on single measures model with a 95% CI. An ICC above .70 is deemed sufficient for group comparison, but for individual monitoring, the ICC should exceed .90 to .95.[39]

Even with a high ICC, the trial-to-trial consistency of physical measurements can be poor, especially in heterogeneous data sets.[20–22] Thus, we also considered their absolute reliability,[21,22] which we calculated with the Bland-Altman 95% limits of agreement (LoA) and SEM.[20,40,41] To facilitate interpretation of the results, the SEM is reported in the same quantity used for the original measurement (eg, kilograms for grip strength, meters per second for speed, seconds for time). It thus provides the range within which a participant's true score may fall.[42] If the SEM is small, indicating high absolute reliability, the true score is close to the recorded score.[20] The probabilities of the normal curve then can be applied to the SEM,[11] meaning 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[20]:

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

The 95% CIs for the SEM were calculated as described by Stratford and Goldsmith[43]:

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

The abbreviations in the latter formula have the following meaning: SSE=the sum of squared errors in the analysis of variance (ANOVA) table; $\chi^2_{\alpha,dfe}$=the chi-square value for probability level α; and dfe=the degrees of freedom of the SSE provided in the ANOVA table.[43] The square roots of these 2 values provide the borders for the 95% CI of the SEM.[43]

Finally, to be able to interpret changes in test scores, the MDC with 95% CI was calculated[11]:

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

The MDC is the required magnitude of observable change that exceeds the anticipated measurement error and within-subject variability.[44] In other words, if a participant's score exceeds the value of the MDC, 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 cognitive impairment (MMSE≥20) and those with moderate cognitive impairment (MMSE=10–19).[37,38] 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.[45]

For a visual inspection of the similarity between the 2 measurements, Bland-Altman plots were created with the LoA. For nonskewed data, the following formula was used to calculate the LoA[46]:

Mean difference ± 1.96 SD.

For skewed data, the following formula was used to calculate the LoA[46]:

$$\pm 2\overline{X}(10^a - 1)/(10^a + 1)$$

$$a = 1.96{*}\sqrt{} \cdot 2\sigma^2_{ER}$$

with σ^2_{ER} reflecting the residual-error variance.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

Table 1.
Characteristics of the Participants[a]

Variable	Total Group (N=58)	Mild Cognitive Impairment (n=30)	Moderate Cognitive Impairment (n=28)	*P*
Age (y)				.89[b]
$\bar{X}$	82.47	82.37	82.57	
SD	5.31	5.16	5.55	
Minimum-maximum	70–91	70–91	70–92	
Sex (female), n (%)	41 (70.7%)	21 (70.0%)	21 (75%)	.49[c]
Place of residence	NH 34% HE 12% HL 53%	NH 23% HE 3% HL 73%	NH 46% HE 21% HL 32%	.01[c]
MMSE				<.001[b]
$\bar{X}$	19.24	22.77	15.46	
SD	4.37	2.13	2.63	
Minimum-maximum	10–28	20–28	10–19	
Walking aid (yes)	26 (44.8%)	15 (50.0%)	11 (39.3%)	.41[c]

[a] NH=nursing home, HE=home for the elderly, HL=home living, MMSE=Mini-Mental State Examination.
[b] *t* test.
[c] Chi-square test.

Role of the Funding Source

The research was funded by the University Medical Center Groningen, Groningen, the Netherlands.

Results

Table 1 presents the characteristics of the 58 participants in the final sample. Seventeen participants were male, and 41 were female, with ages ranging from 70 to 92 years. No significant differences in age, sex, or the use of walking aids were found between the participants with mild cognitive impairment (MMSE=20–28) and those with moderate cognitive impairment (MMSE=10–19). However, the differences for place of residence were statistically significant.

Table 2 presents the relative and absolute reliability values for the 6 physical performance tests for the total group. The relative reliability of the F8W, the TUG, and Jamar dynamometer was excellent (ICC>.90), and good for the 6-m walk test, the CRT, and the FICSIT–4 (ICC=.75–90). The width of the CI of the ICCs ranged between .05 and .20, with the TUG having the smallest CI and the FICSIT–4 having the largest CI. The absolute reliability of the tests, measured with the SEMs and MDCs, was large.

The Figure shows the Bland-Altman plots with the 95% LoA for the 6 tests calculated for the total group.[40,46] The data of the F8W, the TUG, and the 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-m walk test, the CRT, and the FICSIT–4 were

Table 2.
Descriptive and Reliability Measures of the Physical Performance Tests in the Study Group Based on a 1-Week Test-Retest Interval[a]

Measure	n	Test	Retest	KT	F Value	*P*	ICC	CI_{95} ICC	SEM	CI_{95} SEM	MDC_{95}
		$\bar{X}$ (SD) [Minimum-Maximum]	$\bar{X}$ (SD) [Minimum-Maximum]								
6-m walk test (m/s)	58	0.77 (0.25) [0.32–1.60]	0.75 (0.28) [0.24–1.85]	0.42	−0.48	.49	.86	.78–.92	0.10	0.08–0.12	0.27
F8W (s)	46	45.97 (21.23) [19.26–120.00]	45.51 (20.90) [17.30–114.82]	0.00	0.14[b]	.71	.91[b]	.85–.95	6.26	5.41–8.21	17.35
TUG (s)	58	18.55 (9.74) [7.83–67.65]	18.68 (9.01) [8.41–58.50]	0.00	0.99[b]	.32	.94[b]	.92–.97	2.12	1.74–2.52	5.88
FICSIT–4 (points)	58	2.55 (1.10) [0.00–4.00]	2.58 (1.32) [0.00–5.00]	0.84	0.06	.80	.79	.67–.87	0.55	0.47–0.69	1.52
Chair Rise Test (n)	52	8.12 (2.95) [2.00–14.50]	8.30 (3.32) [2.00–18.00]	0.87	0.54	.47	.84	.73–.90	1.26	1.06–1.57	3.49
Jamar dynamometer (kg)	57	20.77 (9.18) [9.00–55.00]	20.55 (8.34) [10.00–46.00]	0.01	0.01[b]	.95	.90[b]	.84–.94	2.74	2.05–2.98	7.59

[a] KT=Koenker test for heteroscedasticity, ICC=intraclass correlation, CI_{95}=95% confidence interval, SEM=standard error of measurement, MDC=minimal detectable change, F8W=Figure-of-Eight Walk Test, TUG=Timed "Up-and-Go" Test, FICSIT–4=Frailty and Injuries: Cooperative Studies of Intervention Techniques–4.
[b] Calculation over log-transferred data.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

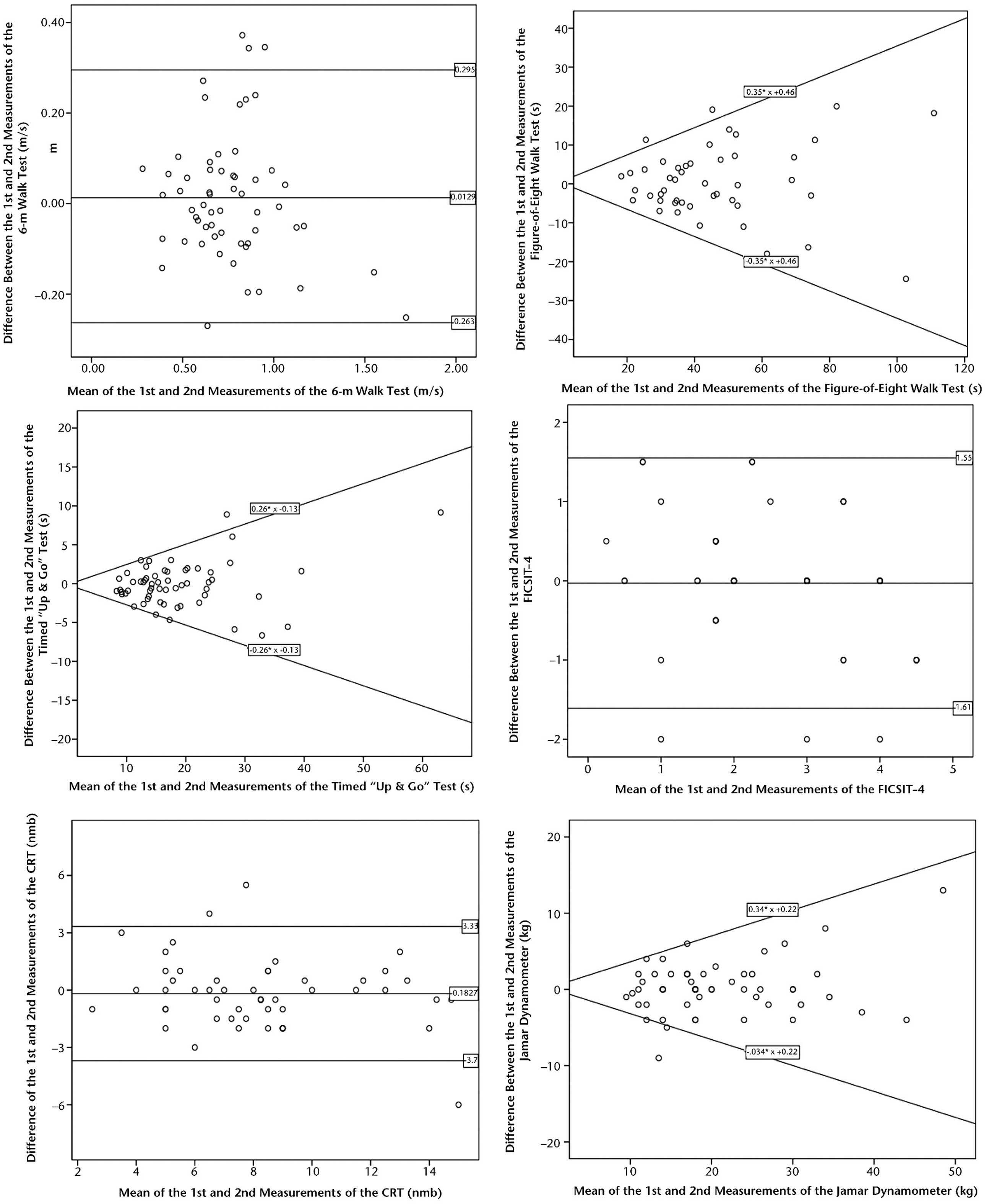

Figure.

Bland-Altman plots showing the levels of agreement for the heteroscedastic and the homoscedastic data for the 6 tests evaluated. The 2 measurements were 1 week apart. CRT=Chair Rise Test, FICSIT–4=Frailty and Injuries: Cooperative Studies of Intervention Techniques–4, nmb=number.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

Table 3.
Baseline and Retest Outcomes (and Standard Deviations) and Reliability Values for the 6 Physical Performance Tests Stratified by Current Cognitive Functioning[a]

Variable	Measure	6-m Walk Test (m/s)	F8W (s)	TUG (s)	FICSIT–4	Chair Rise Test (n)	Jamar Dynamometer (kg)
Mild cognitive impairment (MMSE=20–28)	n	30	25	30	30	29	29
	Session 1, $\bar{X}$ (SD)	0.74 (0.26)	48.12 (25.21)	16.95 (7.49)	2.72 (1.14)	9.12 (3.11)[b]	20.83 (7.87)
	Session 2, $\bar{X}$ (SD)	0.73 (0.30)	45.61 (24.93)	17.01 (6.96)	2.83 (1.29)	9.33 (3.56)[c]	20.97 (6.84)
	ICC (CI_{95})	.83 (.67–.91)	.94 (.86–.97)	.96 (.92–.98)	.82 (.65–.91)	.79 (.60–.90)	.86 (.72–.93)
	SEM (CI_{95})	0.11 (0.09–0.11)	6.24 (5.63–10.03)	1.43 (1.06–1.79)	0.59 (0.48–0.81)	1.52 (1.22–2.08)	2.75 (1.85–3.15)
	MDC_{95}	0.29	17.30	3.96	1.64	4.21	7.62
Moderate cognitive impairment (MMSE=10–19)	n	28	21	28	28	23	28
	Session 1, $\bar{X}$ (SD)	0.80 (0.25)	43.42 (15.41)	20.26 (11.59)	2.38 (1.04)	6.85 (2.21)[b]	20.71 (10.52)
	Session 2, $\bar{X}$ (SD)	0.78 (0.26)	45.51 (15.39)	20.46 (10.63)	2.30 (1.31)	7.00 (2.49)[c]	20.13 (9.77)
	ICC (CI_{95})	0.89 (0.78–0.95)	0.85 (0.67–0.94)	0.94 (0.87–0.97)	0.80 (0.61–0.90)	0.88 (0.73–0.95)	0.94 (0.87–0.97)
	SEM (CI_{95})	0.09 (0.07–0.13)	6.00 (4.01–7.58)	2.91 (2.10–3.61)	0.60 (0.48–0.82)	0.83 (0.65–1.04)	2.57 (2.02–3.47)
	MDC_{95}	0.25	16.63	8.07	1.66	2.30	7.11

[a] F8W=Figure-of-Eight Walk Test, TUG=Timed "Up-and-Go" Test, FICSIT–4=Frailty and Injuries: Cooperative Studies of Intervention Techniques–4, CI_{95}=95% confidence interval.
[b] Significant difference at baseline between participants with higher and lower scores on the MMSE ($P<.01$).
[c] Significant difference at retest between participants with higher and lower scores on the MMSE ($P<.01$).

homoscedastic and, consequently, had a constant LoA.

Table 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 milder cognitive deficits achieving better scores. We found no significant between-group difference for relative reliability, but the absolute reliability 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 seconds) in participants with mild cognitive impairment versus those with moderate cognitive impairment (8.07 seconds). The MDC of the CRT was larger (4.21 stands) in participants with mild cognitive impairment versus those with moderate cognitive impairment (2.30 stands).

Discussion

The main goal of our study was to evaluate the relative and absolute reliability of 6 physical functioning tests in older people (70–92 years) 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, F8W, and Jamar dynamometer (ICC>.90) and good for the 6-m walk test, CRT, and FICSIT–4 (ICC=.75–.90). The differences in relative reliability between the participants with mild cognitive impairment and those with moderate cognitive impairment were nonsignificant.

The values we obtained for the F8W, Jamar dynamometer, 6-m walk test, and CRT were similar to those earlier reported for similarly aged participants with[24] and without[18,31,36] dementia. The values we recorded for the TUG were somewhat lower than those Ries et al reported for patients with Alzheimer disease (ICC=.985–.988).[11] It is likely that this disparity was caused by differences in the characteristics of the 2 patient groups. The percentage of female participants in our sample was higher than that in the study by Ries and colleagues.[11]

A study solely evaluating female patients with different subtypes of dementia showed lower relative reliability scores for the TUG (ICC=.87) and the dynamometer test (ICC=.70).[24] 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 a single group, it causes an upward bias in the reliability coefficient.

The TUG, F8W, and Jamar dynamometer values exceeded the threshold

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

for minimal acceptable reliability (ICC=.90) and thus may be useful for individual monitoring.[11,39] However, for that goal, the absolute reliability also should be considered to establish the within-subject test-retest variability, which we do in the next section.

Given their lower ICC scores, the 6-m walk test, the CRT, and the FICSIT–4 do not appear suitable for individual performance monitoring. However, because all 6 tests exceeded the threshold for group comparisons (ICC>.70),[39] 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.[47] The SEM and the MDC are easy to interpret because they are expressed in the same units as the original measure and, as such, are very useful for clinicians to determine individual improvement.[42] They conveniently allow the 95% CI (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%).[11] Because the MDC and SEM are so closely linked, this discussion will focus solely on the MDC.

To interpret the MDC correctly, the variance of the data should remain constant with increasing means (homoscedastic distribution). A homoscedastic distribution was true for the 6-m walk test, FICSIT–4, and CRT. It required an improvement of 0.27 m/s and an increase of 1.52 points for the MDCs of the 6-m 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 reliability for the participants with moderate cognitive impairment is explained by a floor effect.

For the F8W, the TUG, and Jamar dynamometer, the variance did not remain constant with incremental means (heteroscedastic distribution; see Figure). 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 in the Figure).[46] This finding indicates that the participants who attained lower scores on these 3 tests showed less variability than their peers achieving higher scores. Consequently, for the F8W, 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). These problems should be kept in mind when interpreting their respective MDCs.

For the F8W to exceed the MDC, an improvement of 17.35 seconds was required, and 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 seconds for participants with mild cognitive impairment and 8.07 seconds for those with moderate cognitive impairment. The distinction on the TUG between participants with mild and moderate cognitive impairment is in line with the findings of a study among patients with Alzheimer disease.[11]

Although the MDC should facilitate the appraisal of individual improvement on certain tests, the large margins of improvement the tests appeared to require (eg, 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. The second issue we will address is 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[6] shows that only 1 study out of 16 showed a postintervention improvement larger than the MDCs for the TUG, the sit-to-stand test, and gait and balance abilities measured with the Tinetti scale.[35] This finding suggests that improvements exceeding the MDC are not viable; 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 patients with dementia, van Iersel et al concluded that an increased walking speed of 0.21 m/s reduced the (expert-rated) risk of falling.[48] This value is below the MDC computed in our study (0.27 m/s), rendering gait speed, as measured with the 6-m walk test, a less suitable measure to detect changes of this magnitude in fall risk. The more sophisticated GAITRite (CIR Systems Inc, Sparta, New Jersey) walkway system yielded a smaller MDC (0.11 m/s)[11] 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

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

TUG as clinically relevant.[48] 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 improvements in older people with dementia.[11,49,50]

In summary, we conclude that the MDCs obtained for the 6 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 (eg, Alzheimer 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 (the cutoff for mild cognitive deficit). All 6 participants were attending geriatric adult day care. These findings mean that they had diagnoses of dementia according to the DSM-IV criteria, which is necessary for approval by the CIZ 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 (r=.50), it is not likely that the use of hands had a large effect on the outcome of our CRT, 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 individuals was required, but as 58 participants completed our 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, the examiners could not be completely blinded from the level of cognitive functioning. The examiners did not, however, have any information regarding the MMSE scores of the participants at the moment of testing.

Finally, although the generalizability of our study appears adequate given the heterogeneity of the participants, its generalizability might be hampered by the limited geographical variability.

Conclusion

The relative reliability of the 6 physical performance tests—6-m walk test, F8W, TUG, FICSIT-4, CRT, and the Jamar dynamometer—was 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 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 people with dementia and to define criteria for clinically relevant changes.

All authors provided concept/idea/research design and writing. Mr Blankevoort provided data collection. Dr Scherder provided data analysis. The authors thank all of the students, participants, and institutions for their cooperation.

The research was funded by the University Medical Center Groningen.

DOI: 10.2522/ptj.20110164

References

1 Mura T, Dartigues JF, Berr C. How many dementia cases in France and Europe? Alternative projections and scenarios 2010–2050. *Eur J Neurol*. 2010;17:252–259.

2 Kido T, Tabara Y, Igase M, et al. Postural instability is associated with brain atrophy and cognitive impairment in the elderly: the J-SHIPP study. *Dement Geriatr Cogn Disord*. 2010;29:379–387.

3 Leandri M, Cammisuli S, Cammarata S, et al. Balance features in Alzheimer's disease and amnestic mild cognitive impairment. *J Alzheimers Dis*. 2009;16:113–120.

4 Wennie Huang WN, Perera S, VanSwearingen J, Studenski S. Performance measures predict onset of activity of daily living difficulty in community-dwelling older adults. *J Am Geriatr Soc*. 2010;58:844–852.

5 Iavarone A, Milan G, Vargas G, et al. Role of functional performance in diagnosis of dementia in elderly people with low educational level living in southern Italy. *Aging Clin Exp Res*. 2007;19:104–109.

6 Blankevoort CG, van Heuvelen MJ, Boersma F, et al. Review of effects of physical activity on strength, balance, mobility and ADL performance in elderly subjects with dementia. *Dement Geriatr Cogn Disord*. 2010;30:392–402.

7 Rydwik E, Frandin K, Akner G. Effects of physical training on physical performance in institutionalised elderly patients (70+) with multiple diagnoses. *Age Ageing*. 2004;33:13–23.

8 Kuiack SL, Campbell WW, Evans WJ. A structured resistive training program improves muscle strength and power in elderly persons with dementia. *Act Adapt Aging*. 2004;8:35.

9 Rolland Y, Pillard F, Klapouszczak A, et al. Exercise program for nursing home residents with Alzheimer's disease: a 1-year randomized, controlled trial. *J Am Geriatr Soc*. 2007;55:158–165.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022

10 Podsiadlo D, Richardson S. The timed "up & go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc*. 1991;39:142–148.

11 Ries JD, Echternach JL, Nof L, Gagnon Blodgett M. Test-retest reliability and minimal detectable change scores for the timed "up & go" test, the six-minute walk test, and gait speed in people with Alzheimer disease. *Phys Ther*. 2009;89:569–579.

12 Johannson JO, Jarnlo GB. Balance training in 70-year-old women. *Physiother Theory Pract*. 1991;7:121–125.

13 Shkuratova N, Morris ME, Huxham F. Effects of age on balance control during walking. *Arch Phys Med Rehabil*. 2004;85: 582–588.

14 Rossiter-Fornoff JE, Wolf SL, Wolfson LI, Buchner DM. A cross-sectional validation study of the FICSIT common data base static balance measures. Frailty and Injuries: Cooperative Studies of Intervention Techniques. *J Gerontol A Biol Sci Med Sci*. 1995;50:M291–M297.

15 Resnick B, Gruber-Baldini AL, Zimmerman S, et al. Nursing home resident outcomes from the res-care intervention. *J Am Geriatr Soc*. 2009;57:1156–1165.

16 Mathiowetz V, Weber K, Volland G, Kashman N. Reliability and validity of grip and pinch strength evaluations. *J Hand Surg*. 1984;9:222–226.

17 van Heuvelen MJ, Kempen GI, Brouwer WH, de Greef MH. Physical fitness related to disability in older persons. *Gerontology*. 2000;46:333–341.

18 Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport*. 1999;70:113–119.

19 Baumgartner TA, Jackson AS, Rowe DA, Mahar M. *Measurement for Evaluation in Physical Education and Exercise Science.* Columbus, OH: McGraw-Hill Humanities; 2003:560.

20 Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy*. 2000;2:94–99.

21 Nordin E, Rosendahl E, Lundin-Olsson L. Timed "up & go" test: reliability in older people dependent in activities of daily living—focus on cognitive state. *Phys Ther*. 2006;86:646–655.

22 van Iersel MB, Benraad CE, Rikkert MG. Validity and reliability of quantitative gait analysis in geriatric patients with and without dementia. *J Am Geriatr Soc*. 2007;55: 632–634.

23 Phillips CD, Chu CW, Morris JN, Hawes C. Effects of cognitive impairment on the reliability of geriatric assessments in nursing homes. *J Am Geriatr Soc*. 1993;41:136–142.

24 Thomas VS, Hageman PA. A preliminary study on the reliability of physical performance measures in older day-care center clients with dementia. *Int Psychogeriatr*. 2002;14:17–23.

25 *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed, text rev. Washington, DC: American Psychiatric Association; 2000.

26 Folstein MF, Folstein SE, McHugh PR. "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12: 189–198.

27 Tappen RM, Roach KE, Buchner D, et al. Reliability of physical performance measures in nursing home residents with Alzheimer's disease. *J Gerontol A Biol Sci Med Sci*. 1997;52:M52–M55.

28 Pettersson AF, Engardt M, Wahlund LO. Activity level and balance in subjects with mild Alzheimer's disease. *Dement Geriatr Cogn Disord*. 2002;13:213–216.

29 Frandin K, Sonn U, Svantesson U, Grimby G. Functional balance tests in 76-year-olds in relation to performance, activities of daily living and platform tests. *Scand J Rehabil Med*. 1995;27:231–241.

30 Tegner Y, Lysholm J, Lysholm M, Gillquist J. A performance test to monitor rehabilitation and evaluate anterior cruciate ligament injuries. *Am J Sports Med*. 1986;14: 156–159.

31 Jarnlo GB, Nordell E. Reliability of the modified figure of eight—a balance performance test for elderly women. *Physiother Theory and Pract*. 2003;19:35–43.

32 Helbostad JL, Sletvold O, Moe-Nilssen R. Effects of home exercises and group training on functional abilities in home-dwelling older persons with mobility and balance problems: a randomized study. *Aging Clin Exp Res*. 2004;16:113–121.

33 Eggermont LH, Gavett BE, Volkers KM, et al. Lower-extremity function in cognitively healthy aging, mild cognitive impairment, and Alzheimer's disease. *Arch Phys Med Rehabil*. 2010;91:584–588.

34 Rikli RE, Jones CJ. *Senior Fitness Test Kit—Updated Edition.* Champaign, IL: Human Kinetics; 2007.

35 Santana-Sosa E, Barriopedro MI, Lopez-Mojares LM, et al. Exercise training is beneficial for Alzheimer's patients. *Int J Sports Med*. 2008;29:845–850.

36 Bohannon RW, Schaubert KL. Test-retest reliability of grip-strength measures obtained over a 12-week interval from community-dwelling elders. *J Hand Ther*. 2005;18:426–427, quiz 428.

37 Binetti G, Mega MS, Magni E, et al. Behavioral disorders in Alzheimer disease: a transcultural perspective. *Arch Neurol*. 1998; 55:539–544.

38 Kapaki E, Paraskevas GP. The cognitive effects of cholinesterase inhibitor treatment in every-day practice. *Curr Med Res Opin*. 2005;21:871–875.

39 Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11:193–205.

40 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.

41 Keating J, Matyas T. Unreliable inferences from reliable measurements. *Aust J Physiother*. 1998;44:5–10.

42 Domholdt E. *Rehabilitation Research: Principles and Applications.* 3rd ed. St Louis, MO: Elsevier Saunders; 2005.

43 Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77: 745–750.

44 Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther*. 2006; 86:735–743.

45 Flechner L, Tseng TY. Understanding results: P-values, confidence intervals, and number need to treat. *Indian J Urol*. 2011; 27:532–535.

46 Euser AM, Dekker FW, le Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. *J Clin Epidemiol*. 2008;61:978–982.

47 Overend T, Anderson C, Sawant A, et al. Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiother Can*. 2010; 62:122–128.

48 van Iersel MB, Munneke M, Esselink RA, et al. Gait velocity and the timed-up-and-go test were sensitive to changes in mobility in frail elderly patients. *J Clin Epidemiol*. 2008;61:186–191.

49 van Iersel MB, Hoefsloot W, Munneke M, et al. Systematic review of quantitative clinical gait analysis in patients with dementia. *Z Gerontol Geriatr*. 2004;37: 27–32.

50 Rabheru K. Disease staging and milestones. *Can J Neurol Sci*. 2007;34(suppl 1):S62–S66.

Downloaded from https://academic.oup.com/ptj/article/93/1/69/2735322 by guest on 20 August 2022