Test-Retest Reliability and Minimal Detectable Change on Balance and Ambulation Tests, the 36-Item Short-Form Health Survey, and the Unified Parkinson Disease Rating Scale in People With Parkinsonism

Teresa Steffen, Megan Seney

Background and Purpose. Distinguishing between a clinically significant change and change due to measurement error can be difficult. The purpose of this study was to determine test-retest reliability and minimal detectable change for the Berg Balance Scale (BBS), forward and backward functional reach, the Romberg Test and the Sharpened Romberg Test (SRT) with eyes open and closed, the Activities-specific Balance Confidence (ABC) Scale, the Six-Minute Walk Test (6MWT), comfortable and fast gait speed, the Timed "Up & Go" Test (TUG), the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36), and the Unified Parkinson Disease Rating Scale (UPDRS) in people with parkinsonism.

Subjects. Thirty-seven community-dwelling adults with parkinsonism (mean age=71 years) participated. The Hoehn and Yahr Scale median score of 2 was on the lower end of the scale; however, the scores ranged from 1 to 4.

Methods. Subjects were tested twice by the same raters, with 1 week between tests. Test-retest reliability was calculated using intraclass correlation coefficients (ICCs). Minimal detectable change was calculated using a 95% confidence interval (MDC_{95}).

Results. The ICCs for test-retest reliability were above .90 for the BBS, ABC Scale, SRT with eyes closed, 6MWT, and comfortable and fast gait speeds. The MDC_{95} values for those functional tests were: BBS=5/56, ABC Scale=13%, SRT with eyes closed=19 seconds, 6MWT=82 m, comfortable gait speed=0.18 m/s, and fast gait speed=0.25 m/s. The ICCs for test-retest reliability of SF-36 scores were above .80, with the exception of the social functioning subscale. The MDC_{95} values for the SF-36 ranged between 19% and 45%. The MDC_{95} values for the UPDRS Activities of Daily Living section, Motor Examination section, and total scores were 4/52, 11/108, and 13/176, respectively.
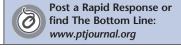
Discussion and Conclusion. Minimal detectable change values are useful to therapists in rehabilitation and wellness programs in determining whether change during or after intervention is clinically significant. High test-retest reliability of scores for the BBS, ABC Scale, SRT with eyes closed, 6MWT, and gait speed make them trustworthy functional assessments in people with parkinsonism. The SF-36 and UPDRS provide quality-of-life and disease severity rating values in the ongoing assessment of people with parkinsonism.

T Steffen, PT, PhD, is Professor, Program in Physical Therapy, Concordia University Wisconsin, Mequon, Wis. Dr Steffen's mailing address is: N102 W7525 St James Ct, Cedarburg, WI 53102 (USA). Address all correspondence to Dr Steffen at: csteffen1@wi.rr.com.

M Seney is a student in the Program in Physical Therapy, Concordia University Wisconsin.

Post a Rapid Response or find The Bottom Line: www.ptjournal.org

Physical therapists strive to create interventions that focus on improving a patient's functional ability. Function gained during or after therapy often is measured by change in scores on a functional assessment instrument over time.[1] When results improve from one assessment to another, therapists often assume that the patient has progressed. Unfortunately, there is a chance the difference between assessments is a result of measurement error.[2] A common problem involves deciding whether the results are clinically significant or an error in measurement. To determine whether an improvement is significant, researchers need to provide minimal detectable change (MDC) scores, by patient population, for tests. *Minimal detectable change* is defined as the minimal amount of change that is not due to variation in measurement.[3]

Clinicians can interpret MDC scores as the minimal change that is not due to error. Scores at or above the MDC level are due to patient improvement on the test rather than measurement error. Measurement error includes expected or typical variability in patient performance. In the literature, various methods are utilized to calculate change scores, including the standard error of measurement (SEM), minimal clinically important difference (MCID), and smallest detectable difference (SDD). The SEM is calculated by multiplying the standard deviation by $\sqrt{1}$ minus the reliability coefficient, which is the stability or variability of response and indicates the range of the scores that can be expected upon retesting.[4] The MCID is the smallest meaningful change, as judged by the patient or experts in the field,[3] and is determined by questioning or observing the patient. Some researchers refer to the MDC utilizing a 95% confidence interval (MDC_{95}) as the SDD.[5]

Once the MDC is determined on a particular test for a given population, therapists can interpret whether the change score for their patient is at or above the minimal level of detectable change reported in the literature. If the patient's score is less than the MDC value, it is considered to be indistinguishable from measurement error. Accordingly, a patient who demonstrates less than the MDC value is viewed as not benefiting from the intervention. For example, following hip fracture, the MDC is 0.08 m/s for comfortable gait speed.[6] If a patient's comfortable gait speed increases less than 0.08 m/s, the change is within measurement error, leading to the conclusion that a clinically significant change did not occur as a result of the therapeutic intervention.

To evaluate MDC, researchers first must measure test-retest reliability. On functional tests, a 7-day separation period typically is used. Sources of error may include inconsistencies caused by the participant's physical or mental condition, variations in the testing procedure, or tester error. Maintaining consistency and using standardized protocols for testing, such as using the same tester, setup, testing order, and time of day, can improve test-retest reliability.

The MDC is based on the SEM and is calculated using the following formula[3]:

$$MDC = z\text{-score}_{\text{level of confidence}} \times SD_{\text{baseline}} \times \sqrt{(2[1 - r_{\text{test-retest}}])}.$$

The z-score represents the confidence interval from a normal distribution, SD is the standard deviation at baseline, and r is the test-retest reliability coefficient. The multiplier of $\sqrt{2}$ is used to account for the additional uncertainty introduced by using difference scores from measurements at 2 points in time. Some researchers[1,3] suggest using a confidence interval of 90% due to its use being more common in the literature; however, a confidence interval of 95% increases the precision of score estimation and is the SDD.[3] Internal consistency, determined by the Cronbach alpha, of a multiple-item test such as the Berg Balance Scale (BBS), the Activities-specific Balance Confidence (ABC) Scale, the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36), and the Unified Parkinson Disease Rating Scale (UPDRS) sometimes replaces r or intraclass correlation coefficients (ICCs) in the MDC formula when test-retest reliability is not reported. Internal consistency is the extent to which multiple items within a scale or subscale measure one characteristic and nothing else. Internal consistency of a multiple-item test is considered high if it approaches a Cronbach alpha of .90 in a given population. Test-retest scores are considered a more conservative approach when calculating MDC values[3] in situations where both internal consistency and test-retest reliability are reported. Test-retest reliability and internal consistency reliability are necessary forms of reliability that should be reported in multiple-item tests in which item scores are summed or averaged.

One patient population for which MDC scores would be useful in helping to distinguish actual change from measurement error is people with parkinsonism. Parkinsonism is a constellation of symptoms, including tremor, rigidity, bradykinesia (slow movements), and loss of postural reflexes. Although Parkinson disease (PD) is the most frequent cause of parkinsonism, it includes other diagnoses such as Parkinson-plus syndromes of progressive supranuclear palsy and corticobasal degeneration. Parkinsonism symptoms create functional limitations of balance—which often are measured in the clinic with the BBS, ABC Scale, Func-

tional Reach Test (FRT), Romberg Test (RT), and Sharpened Romberg Test (SRT)—and difficulties in mobility—which often are measured with the Six-Minute Walk Test (6MWT), Timed "Up & Go" Test (TUG), and gait speed. The disease also affects quality of life (measured most often with the SF-36). Disease severity is measured in people with PD with the UPDRS.

Literature Review of MDC Values

Extensive literature searches were done to find all previous test-retest articles published up to March 2007 on each of the instruments listed above. This was done to determine whether MDCs could be calculated for each of the instruments. When test-retest ICCs or Pearson (r) values and standard deviations were reported in the literature, those values were used to calculate MDC_{95} for that study.

BBS

The BBS is a 14-item test, using ordinal scoring from 0 to 4 for each item, designed to measure static and dynamic standing balance. The total score range is 0 to 56, with higher scores indicating better balance. The internal consistency of the BBS is moderate to high, ranging from .85 to .98.[7-12] In 3 studies,[9,13,14] ICCs for test-retest reliability of .97 to .99 were reported in subjects with stroke and traumatic brain injury (TBI), respectively. One study performed over 1 week on 26 people with PD reported an ICC of .87.[15] Four studies contained sufficient data to determine MDC_{95}. The MDC_{95} values were 2 for 26 people with PD,[15] 5 for 24 elderly people with or without stroke,[9] 3 for 20 people with hemiparesis,[13] and 4 for 5 people with TBI.[14] All of these studies were performed 7 days apart. The high test-retest reliability, moderate to high internal consistency, and low MDC_{95} scores in these stud-

ies indicate the BBS is a valuable measure to monitor responsiveness to change in patients with neurological disease.

FRT

The FRT is a static balance test designed to measure margins of stability. Based on a review of 10 articles, test-retest reliability for functional reach has been shown to vary from low to high, with ICCs ranging from .42 to .93. Nine studies[15-23] examined forward functional reach, and 1 study[19] examined backward functional reach, with the time between tests varying greatly from 1 day to 1 month.[15-23] Only 3 studies examining test-retest reliability had a sample size over 30.[16,17,24] Three studies[15,20,22] reported test-retest reliability in subjects with PD. One study of 26 subjects with idiopathic PD reported an ICC of .74 for the FRT with a testing interval of 1 week,[15] and a study of 14 subjects with PD reported an ICC of .84 for the FRT with a testing interval of 1 day.[20] Another study of 10 elderly subjects with no known neurological impairment and 20 subjects with PD, using a testing interval of a week, reported ICC (2,1) values of .62 for the subjects with no known neurological impairment, .93 for subjects with PD who had a history of falls, and .42 for subjects with PD with no history of falls.[22]

Of the current studies examining test-retest reliability of the FRT, 4 studies[19,21-23] provided enough data to calculate MDC_{95}, which ranged from 4 to 11 cm. Two studies reporting test-retest reliability of the FRT in 20 people with PD, with tests 1 week apart, demonstrated MDC_{95} values of 4 cm for people who had fallen, 8 cm for people who had not fallen, and 12 cm for 26 people with a diagnosis of idiopathic PD.[15,22] Studies on forward functional reach have provided a wide range of MDC_{95} values for people with PD,

and no MDC_{95} data on backward functional reach.

RT and SRT

The RT and SRT are tests of static balance that measure the ability to maintain balance or equilibrium with a narrowed base of support. Currently, there are no studies that have examined test-retest reliability of the RT or SRT for subjects with PD. In one study of 30 subjects with unilateral vestibular loss, aged 29 to 78 years, test-retest reliability values (ICC [2,2]) were .63 for the SRT with eyes open and .76 for the SRT with eyes closed.[25] In a study of 20 subjects with central neurological dysfunction, aged 58 to 85 years, test-retest reliability values (ICC [2,2]) were .75 for the SRT with eyes open and .97 for the SRT with eyes closed.[25] In 2 studies, one with 18 volunteers aged 24 to 39 years[17] and one with 45 volunteers aged 55 to 75 years,[26] test-retest reliability values (ICC [2,1][17] and r[26]) were .72 and .76 for the SRT with eyes closed and .90 for the SRT with eyes open. One study with a small sample size (n=12) used the coefficient of variation (CV), which indicates an association between 2 variables, and showed a high degree of variability, ranging from .14 to .86, between the tests on 5 consecutive days.[27] In 2 studies, MDC_{95} scores for the SRT with eyes open ranged from 9 to 10 seconds,[25] and MDC_{95} scores for the SRT with eyes closed ranged from 3 to 9 seconds.[25,26] Test-retest studies are needed on the RT and SRT for populations with neurological disorders, including people with PD.

ABC Scale

The ABC Scale is a 16-item questionnaire used to measure balance confidence in specific situations, with scores ranging from 0% to 100%. Internal consistency of the ABC Scale in 4 studies[19,28-30] ranged from .80 to .98. These 4 studies also addressed test-retest reliability of the

ABC Scale. The time between testing dates ranged from 1 to 4 weeks in various populations, including personal care home residents, patients from outpatient clinics, and community-dwelling older adults. The test-retest reliability values (ICC [1,1],[19] ICC,[28] ICC[2,1],[29] and r[30]) were .70 to .92.[19,28-30] Two of the 4 studies[28,29] had sample sizes greater than 30 subjects, and 3 studies[19,29,30] provided enough data to calculate MDC_{95} scores of 18% to 38%. In these studies,[19,29,30] MDC_{95} values also were calculated using the Cronbach alpha, with results ranging from 6% to 15%. Minimal detectable change values by patient diagnoses, including PD, are needed for the ABC Scale.

6MWT

The 6MWT tests endurance by measuring the maximum distance that a person can walk in 6 minutes. For some patients, it is a submaximal test of aerobic capacity. Eighteen articles were obtained on the test-retest reliability of the 6MWT.[31-48] The majority of these studies used a 7-day interval between tests. None of the 18 studies evaluated test-retest reliability for individuals with PD. Two reliability studies of individuals with stroke demonstrated high ICCs of .99.[31,32] In 2 studies, MDC_{95} values were 34 m without a report of days between tests[31] and 36 m with a 7-day testing interval.[32]

Five studies[33-37] assessed test-retest reliability for subjects with other neurological disorders, with ICCs ranging from .93 to .96. Sample sizes ranged from 12 to 25 subjects, and testing was conducted between 1 and 14 days apart. The MDC_{95} values were 20 m for subjects with chronic poliomyelitis,[35] 53 and 65 m for adults with cerebral palsy,[34] 71 m for subjects with acquired brain injury,[33] and 106 m for subjects with multiple sclerosis.[36]

Five studies[38-42] evaluated test-retest reliability of individuals with cardiac problems, with ICCs ranging from .88 to .97. The MDC_{95} values were 18 m for subjects with congestive heart failure (CHF),[39] 50 m for subjects with peripheral arterial occlusive disease,[41] 51 m for subjects with cardiac rehabilitation,[38] and 74, 86, and 90 m for subjects with heart failure.[40] One study[42] reported an SEM of 15 m for patients with CHF. Standard deviations were variable, sample sizes ranged from 43 to 786 subjects, and days between tests ranged from 1 to 14.

Four studies examined test-retest reliability for subjects with lung disease, with ICCs ranging from .88 to .95.[39,43,44] The MDC_{95} values were 53 and 63 m for subjects with chronic obstructive pulmonary disease (COPD),[43,45] 87 m for subjects with emphysema,[44] and 18 m for subjects with lung disease.[39] One study[46] reported an MCID of 54 m for subjects with COPD. Standard deviations were variable, sample sizes ranged from 15 to 470 subjects, and days between tests ranged from 1 to 10.

Two studies[47,48] assessed test-retest reliability in older adults, with ICCs of .87 and .93. The MDC values were 77 m for community-dwelling elderly subjects,[48] 89 m for those living in retirement homes,[47] and 94 m for those living in community centers.[47] Sample sizes ranged from 5 to 22 subjects, and days between tests ranged from 7 to 14. Overall, the 6MWT is a reliable test, with different MDC_{95} values by client population. Minimal detectable change studies with larger samples are needed on the 6MWT for people with neurological diseases, including those with PD.

TUG

The TUG is a mobility test for the geriatric population. It includes a sit-to-stand component as well as walking 3 m, turning, and returning to the chair. Three studies[49-51] reported test-retest reliability ranging from .92 to .99 in older adults with arthritis in nursing homes or with multiple conditions. The MDC_{95} in one study of 78 participants was calculated as 15 seconds, with this large value being attributed to a large standard deviation of 17 seconds.[50] A study of 9 men with PD, in Hoehn and Yahr (H&Y) stages 3 to 4 and tested over 7 days, had a test-retest reliability value of .75 and an MDC_{95} of 5 seconds.[52] A study of 26 participants with idiopathic PD, tested over 7 days, reported an ICC of .88 and an MDC_{95} of 2 seconds.[15] Studies with larger samples are needed to verify that a change of 2 to 5 seconds is a meaningful change score on the TUG for people in all stages of PD.

Gait Speed

Gait speed is a measure of overall walking performance, but does not include an endurance component. Both fast and comfortable gait speeds often are measured to ensure that patients have the ability to change walking speed. Fifteen articles[13,15,21,32,37,53-62] were found on test-retest reliability of measurements for gait speed. Of these 15 studies, 6 studies[21,32,37,57,58,60] also measured fast gait speed. The MDC_{95} scores were calculated from previous literature, using measured distances from 3.3 to 10 m. One study[15] evaluated test-retest reliability in individuals with PD and reported an ICC of .81 and an MDC_{95} value of 0.19 m/s. Six studies assessed test-retest reliability in individuals with stroke[13,32,53-55] or TBI,[37] with ICC (3,2) values ranging from .94 to .98 for comfortable gait speed.[13,32,37,53,55] One study[37] assessed test-retest reliability of measurements of fast gait speed in people with TBI and reported an ICC of .96. The MDC_{95} values were 0.11 to 0.24 m/s for

comfortable gait speed[32,53,55] and 0.24 m/s for fast gait speed.[32]

Four studies[56-59] examined test-retest reliability in community-dwelling elderly people. Intraclass correlation coefficients ranged from .79 to .95 overall,[56-59] and ICCs for fast gait speed ranged from .87 to .97.[57,58] The MDC_{95} values were 0.25 to 0.29 m/s for comfortable gait speed and 0.25 m/s for fast gait speed in individuals with or without assistive devices.[57] In one study,[58] MDC_{95} values of 0.06 to 0.14 m/s for comfortable gait speed and 0.08 to 0.15 m/s for fast gait speed were calculated for four 10-year cohort groups over the age of 60 years.

Three studies assessed test-retest reliability in patients with musculoskeletal problems, including osteoporosis,[60] knee osteoarthritis,[61] and hip fracture.[21] Intraclass correlation coefficients ranged from .88 to .97 for comfortable gait speed[60-62] and from .91 to .94 for fast gait speed.[21,60] The MDC_{95} values for comfortable and fast gait speeds were 0.25 and 0.30 m/s, respectively, for people with osteoporosis[60] and 0.49 and 0.51 m/s for people with hip fracture.[21] Overall, the MDC_{95} values for both comfortable and fast gait speeds appeared to be about 0.25 m/s or less for populations tested to date, except for patients with hip fracture, with an MDC_{95} value of approximately 0.50 m/s.

SF-36

The SF-36 is a quality-of-life questionnaire developed as a part of the Medical Outcomes Study to assess 8 physical and mental health concepts as seen from the respondent's point of view. These concepts are: (1) limitations in physical activities because of health problems (Physical Functioning), (2) limitations in social activities because of physical or emotional problems (Social Function-

ing), (3) limitations in usual role activities because of physical health problems (Role–Physical), (4) bodily pain (Bodily Pain), (5) psychological distress and well-being (Mental Health), (6) limitations in usual role activities because of emotional problems (Role–Emotional), (7) energy and fatigue (Vitality), and (8) general health perceptions (General Health). These 8 domains are relevant to general functional status and well-being. The survey was designed for self-administration by people 14 years of age and older or for administration by a trained interviewer in person or by telephone. For each scale, item scores are coded, summed, and transformed, with final values (expressed as a percentage) ranging from 0 (worst health) to 100 (best health).

No articles were found on the test-retest reliability of SF-36 scores in patients with PD; however, 17 articles were found for other populations. These articles studied patients with vestibular dysfunction[63]; veterans[64,65]; adults with an intensive care unit stay of greater than 24 to 48 hours[66]; patients with spinal cord injury[67]; patients with confirmed or suspected ischemic stroke[68]; patients with rheumatoid arthritis[69]; patients with systemic lupus erythematosus[70]; patients with knee disorders[71]; elderly patients[72]; patients with ulcerative colitis[73]; patients with low back pain, menorrhagia, suspected peptic ulcer, or varicose veins[73]; a nonclinical normative sample[74]; and general populations in China,[75] the Basque region of Spain,[76] the Hunter region of New South Wales, Australia,[77] Japan,[78] and Sheffield, United Kingdom.[79] The time interval between tests in these studies often was 2 weeks.[64,65,67,69,71,73,75,76,78,79] Other time intervals used were 4 weeks,[72] 3 weeks,[68] 1 week,[66,70,73,74,77] and 2 days.[63,71] Fourteen studies[63-65,68-76,78,79] were self-

administered with either paper or a computer, 2 studies[66,67] were administered via telephone or personal interview, and 1 study[77] was administered via self-administration and telephone interview.

Internal consistency was reported for some of the studies, and the values ranged from .84 to .98 for Physical Functioning, from .83 to .98 for Role–Physical, from .79 to .96 for Bodily Pain, from .72 to .95 for General Health, from .66 to .96 for Vitality, from .39 to .98 for Social Functioning, from .78 to .99 for Role–Emotional, and from .72 to .95 for Mental Health.[65-68,70,72,73,75-79]

The test-retest ICCs, Pearson (r) values, or Spearman (r) values reported in these studies ranged from .34 to .98 for Physical Functioning, from .36 to .97 for Role–Physical, from .35 to .95 for Bodily Pain, from .41 to .93 for General Health, from .36 to .93 for Vitality, from .05 to .96 for Social Functioning, from .23 to .99 for Role–Emotional, and from .30 to .95 for Mental Health.[63-74,76-79]

The ranges in MDC_{95} values were reported, with values ranging from 11 to 63 for Physical Functioning, from 23 to 81 for Role–Physical, from 19 to 54 for Bodily Pain, from 14 to 43 for General Health, from 13 to 49 for Vitality, from 16 to 85 for Social Functioning, from 11 to 110 for Role–Emotional, and from 12 to 57 for Mental Health.[63-67,69-71,73,74,76,77] These large MDC_{95} values occurred because of large standard deviations within groups and the large range of test-retest reliability values among studies. Considering that people with PD often are treated over long periods of time, MDCs should be developed on quality of life (using SF-36 scales), with progress measured at regular intervals.

UPDRS

The UPDRS is the gold standard instrument used to measure disease severity in PD. It has 3 subscales: I—Mentation, Behavior, and Mood (range=0–16), II—Activities of Daily Living (ADL) (range=0–52), and III—Motor Examination (range=0–108). A total score (range=0–176) can be derived by summating the 3 subscales. Lower scores indicate a less involved disease process. The UPDRS has moderate internal consistency values across multiple studies in the 3 subscales and total score. A Cronbach alpha value of .79 has been reported for the Mentation, Behavior, and Mood subscale,[80] Cronbach alpha values of .85 to .92 have been reported for the ADL subscale,[80–83] Cronbach alpha values of .88 to .95 have been reported for the Motor Examination subscale,[80,81,83–85] and a Cronbach alpha value of .96 has been reported for the total UPDRS score.[86]

One study[87] examined the test-retest reliability of UPDRS scores in 400 patients with early stage, mild PD who were not taking medications. The subjects were examined on 2 occasions, separated by an average of 15 days (SD=8). The ICC (1,1) values were .74 for the Mentation, Behavior, and Mood subscale, .85 for the ADL subscale, .90 for the Motor Examination subscale, and .92 for the total score. The calculated MDC_{95} values were 2, 4, 7, and 9, respectively.[87] In 26 ambulatory subjects with idiopathic PD and no comorbidities, test-retest reliability with a 7-day interval between tests was .84 for the Motor Examination subscale and .74 for the total score.[15] The MDCs were 13 and 15, respectively.[15] Test-retest reliability of Motor Examination subscale scores was evaluated in 34 patients with advanced PD on 2 separate occasions, 1 to 3 weeks apart, with an ICC (3,1) of .90.[88]

Minimal detectable changes of 1 to 2 points for the Mentation, Behavior, and Mood subscale, 2 to 4 for the ADL subscale, 7 to 8 for the Motor Examination subscale, and 9 for the total UPDRS encompass the existing studies. Different versions of the UPDRS are being used in studies, and a shorter version is being developed. The lower reliability of the Mentation, Behavior, and Mood subscale scores suggests the need for caution when using reliability values to calculate an MDC value. Physical therapists are most interested in the Motor Examination subscale of the UPDRS to measure responsiveness to change.

The purpose of this study was to determine the MDC_{95} for people with parkinsonism on the following tests and measures: BBS, forward and backward functional reach, RT and SRT, ABC Scale, 6MWT, comfortable and fast gait speeds, TUG, the 8 subscales of the SF-36, and UPDRS (Mentation, Behavior, and Mood subscale, ADL subscale, Motor Examination subscale, and total score).

Method

Subjects

Participants were recruited via bulletin advertisements and flyer distribution at local fitness centers, physical therapy sites, meal sites throughout southeast Wisconsin, Wisconsin PD organizations, church bulletins, newspapers, and other local news media. Previous research study and pro bono clinic participants also were contacted, and advertisements were placed on the Concordia University Wisconsin Web site and in faculty bulletins.

Eligibility for the study was determined by the presence of a clinical diagnosis of PD or Parkinson-plus syndrome. All potential volunteers were contacted by telephone and given an oral questionnaire. Participants were included if they were

able to stand independently for 1 minute and could walk independently with or without the use of an assistive device. Participants were excluded if they reported a history of a heart condition limiting their activity level, experienced a fall as a result of dizziness or fainting within the previous 2 months, or required help with following directions.

A demographic questionnaire (sex, date of birth, date of diagnosis with PD or Parkinson-plus syndrome, ethnicity, living situation, history of falling, other medical conditions, and current medications) was completed on the first day of testing and reviewed in the participant's presence with a researcher to ensure accuracy. Participants were reminded not to change medications during their scheduled test week and to take medications at the same time on both testing days.

During the spring of 2007, 37 participants with PD (n=35) or Parkinson-plus syndromes (n=2) met all inclusion criteria and consented to participate in this study. This sample reflected the general demographics of the PD population, with more men (n=26) than women (n=11) and an elderly age distribution (mean age=71 years, SD=12). There was a wide range of UPDRS total scores (mean=33/176, range=7–70), demonstrating a sample that captured a wide spectrum of disease severity. The average H&Y score was 2 (range=1–4). Distribution of H&Y stages were: 13 subjects in stage 1, 7 subjects in stage 2, 9 subjects in stage 3, 8 subjects in stage 4, and no subjects in stage 5.

The average disease duration was 14 years (SD=6), and participants were primarily of white/non-Hispanic descent (n=36), with 1 participant of Asian/Pacific descent. Of the 37 participants, 32 were living at home with another person, 3 were living at

home alone, and 2 were in assisted living facilities. The mean number of falls in the previous 6 months was 7 (range=0–182); 21 participants had experienced more than one fall. None of the participants changed medications during their testing week, and all participants reported taking medications at the same time on both testing days. Thirty-one participants were using levodopa, with an average of 412 mg/d (SD=310, range=125–1,150). Participants, on average, had 3 comorbidities (SD=2, range=0–6), including 17 with arthritis, 3 with asthma, 7 with a history of cancer, 11 with high blood pressure, 5 with low blood pressure, 1 with diabetes, 8 with a previous fracture, 9 with depression or other mental health condition, 6 with a history of heart disease, 7 with osteoporosis, 1 with stroke, and 14 with other, unspecified comorbidities.

Procedure

Testing was administered at Concordia University Wisconsin. Any classes scheduled to occur in the vicinity of the testing area were relocated to limit interruptions, and barriers were placed to ensure participant privacy. After signing consent forms, testing began with the SF-36 questionnaire and completion of the demographic information. Balance testing followed and consisted of 4 tests administered to each participant in the following order: BBS, forward and backward functional reach, RT and SRT (eyes open and eyes closed), and ABC Scale. The ambulation tests and the UPDRS were administered last and done in the following order: 6MWT, UPDRS, TUG, and comfortable and fast gait speeds. Each day total testing time per participant was approximately 1 hour. Prior to testing, all researchers were trained in their assigned test, and they performed the same duties on each testing day. Researchers who collected the reliability data were monitored by the coinvestigators before and during the testing procedures to maintain accuracy. All researchers had previous patient experience using the functional tests. If an assistive device was used, the type was documented and the participant was required to use it on the subsequent testing day. Thirty-nine participants were scheduled for the study; 2 participants cancelled due to weather or transportation issues. Researchers did not have access to the previous test results on the second day of testing.

Test-retest reliability was established over a period of 7 days in all participants, with the exception of 1 participant, who was tested 10 days apart. Although a 14-day separation may be preferred for the SF-36 questionnaire, a 7-day interval was used based on previous test-retest studies of the other functional assessments.

Balance testing. The method for the BBS test followed the original design,[7] which consists of 14 items scored on a scale of 0 to 4. A score of 0 indicates the participant was unable to complete the task, and a score of 4 indicates the participant was able to complete the task based on the assigned criteria. The floor-to-seat height of the chair used on items 1, 3, 4, and 5 was 47 cm. The height of the chair without armrests used on item 5 was 44.5 cm, and the height of the step stool used on item 12 was 23 cm. A 1.27-cm (0.5-in) slipper was used on item 9. The participants were asked to perform each of the items on the original BBS, with rests as needed. The 14 items were scored by a total of 3 researchers. One researcher scored item 8 while the participants performed the FRT, another researcher scored items 7 and 13 while the participants performed the RT and SRT, and 1 researcher scored the remaining 11 items.

Equipment used for forward and backward functional reach included a level with attached wooden sliders fixed to an adjustable tripod with C rings. Participants were asked to make a fist, raise their dominant arm parallel to the floor with the elbow fully extended, and reach as far forward or backward as possible without losing their balance, lifting their feet off the ground, or touching the equipment. The foot placement and method of reach were not controlled, except to keep the arm at the height of the level. Participants who inquired about foot placement were instructed to stand in a comfortable position. Participants were allowed multiple practice trials until they performed the test correctly. Once a participant was able to perform the test correctly, 2 graded trials were completed. The dominant arm was recorded on the first testing day and used on the second testing day to maintain consistency. The averages of the 2 trials for each direction were used for data analysis, due to the high intratrial reliability reported in previous studies.[10,89] Measurements were recorded (in centimeters) using the third metacarpal as the reference point. Two researchers participated in the data collection. One researcher gave instructions and maintained participant safety. The other researcher adjusted the equipment to match the participant's acromion height, adjusted the wooden slider during reach, and recorded initial and final measurements.

The RT was performed with feet together and eyes open for 60 seconds and with feet together and eyes closed for 60 seconds. The SRT was performed in a tandem standing position, with the dominant foot behind the nondominant foot for 60 seconds with eyes open and for 60 seconds with eyes closed. Timing started after the participant assumed the proper position and stopped if the participant moved his or her feet

from the proper position, touched the table, or opened his or her eyes on the eyes-closed trials or when the maximum balance time of 60 seconds was reached. Participants were given assistance to assume the test position and allowed rest breaks if needed. Up to 3 trials were performed if the maximum balance time was not reached in either of the first 2 trials. Data analysis utilized the longest balance time of all of the trials. Upper-extremity use was not controlled during testing. One researcher administered the RT and SRT to all participants, while another researcher supervised participant safety.

The ABC Scale was administered as an interview consisting of 16 items describing various activities for which participants are asked to rate their confidence in maintaining balance on a scale of 0% (not confident) to 100% (completely confident). Final scores were determined by calculating the average score on the 16 items. To assist participants, an enlarged version of a 0–100 scale was provided.

Ambulation. All ambulation tests were performed on a level tile floor under quiet conditions. The 6MWT was conducted in a 3-m-wide hallway with a 15-m area marked off at 1-m intervals and large cones placed at each end. Participants were read the following instructions: "When I say 'go,' I want you to walk around this track. Keep walking until I say 'stop' or until you are too tired to go any further. If you need to rest, you can stop until you're ready to go again. I am interested in measuring how far you can walk. You can begin when I say 'go.'" The following encouragements were provided: (1) after 1 minute, "You are doing well. You have 5 minutes to go."; (2) at 2 minutes, "Keep up the good work. You have 4 minutes to go."; (3) at 4 minutes, "Keep up the good work.

You have 2 minutes left."; and (4) at 5 minutes, "You are doing well. You have only 1 minute to go." Fifteen seconds prior to completion, participants were informed that time would stop shortly, and the test was stopped at 6 minutes.[90] Total distance walked was measured to the nearest meter.

For the TUG, participants were instructed to sit with their back against a chair (47 cm from floor to seat with armrests), feet behind the tape marker, and arms resting in their lap. Participants were instructed to independently rise on the word "go," comfortably walk a clearly marked distance of 3 m, turn around a cone, walk back to the chair, and sit down with their back against the chair. Time started once the participant's back left the chair and ended when the participant's back returned to the chair. Time to complete the course was measured to the nearest 100th of a second. One practice trial and 2 timed trials were performed; the 2 timed trials were averaged for data analysis.

For the test of comfortable gait speed, participants were asked to walk 10 m and were instructed to "walk at your own comfortable walking speed and stop when you reach the far line." For the test of fast gait speed, participants walked the 10 m with the instructions to "walk as fast as you can safely walk" and to stop at the far line. Time to complete the central 6 m was measured to the nearest 100th of a second using a stopwatch. Time started when any part of the foot crossed the plane of the first tapeline and ended when any part of the foot crossed the plane of the 6-m mark. Rest breaks were allowed between tests or trials, if needed. Participants completed 2 comfortable gait speed trials, followed by 2 fast gait speed trials. The 2 trials were averaged for data anal-

ysis, and gait speeds were calculated (in meters per second).

SF-36. The SF-36 was administered via personal interview by 2 researchers using the interview script provided in the *SF-36 Health Survey: Manual and Interpretation Guide*.[91] Standard procedures for repeating questions and response choices were followed, as outlined in the *SF-36 Health Survey: Manual and Interpretation Guide*.[91] The participants were able to choose from a typed list of response choices that was enlarged and placed on a table in front of them for each question. To avoid influencing the participants' answers on the SF-36, it was the first test given to each participant on both testing days, before they were asked any other health-related questions.

UPDRS. The UPDRS subscales were administered as described by Goetz and colleagues,[92] and a UPDRS total score was calculated based on the sum of the scores of the 3 subscales. The test was administered by 1 of 2 researchers, both of whom reviewed the UPDRS teaching videotape. The original 5-point (1–5) H&Y Scale staging of PD was used in the study.[93] Higher scores on the H&Y Scale indicate greater impairment of PD.

Data Analysis

Internal consistency and test-retest reliability were calculated using SPSS (version 15.0) software.* Internal consistency, assessed using the Cronbach alpha, was calculated for multiple-item tests, such as the BBS, ABC Scale, SF-36, and UPDRS. Internal consistency of .70 or greater was required on the multiple-item test before other forms of reliability were considered trustworthy. The ICC (3,k) was used instead of the Pearson

---

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

correlation coefficient (*r*) for test-retest reliability because it assesses rating reliability by comparing the variability of different ratings of the same subject with the total variation across all ratings and all subjects. For test-retest reliability, either a type 3,1 or type 3,2 ICC was used. The ICC (3,1) was used for the BBS, RT, SRT, ABC Scale, 6MWT, SF-36, and UPDRS because final scores on these tests were based on a single measure from one rater. The ICC (3,2) was used for the TUG, forward and backward functional reach, and comfortable and fast gait speeds because final scores for these tests were based on an average of 2 trials. Normal distribution was assessed for each outcome variable at test day 1 using a histogram plot. Data from 2 participants on the SF-36 and 1 participant on the ABC Scale were excluded from the data analysis due to the presence of cognitive deficits, as judged by the researchers administering the tests. Due to fatigue, the gait speed tests were not administered to one participant.

Table 1.

Internal Consistency for Balance Tests, a Quality-of-Life Measure, and a Disease Severity Rating Scale in People With Parkinsonism (n=36–37)

Test	Internal Consistency		
	Cronbach α for the Present Study		Values From Previous Studies
	Day 1	Day 2	
Balance tests			
Berg Balance Scale	.86	.87	.85–.89[7–12]
Activities-specific Balance Confidence Scale	.95	.96	.80–.98[19,28–30]
36-Item Short-Form Health Survey (SF-36)			
Physical Functioning	.85	.87	.84–.98[65–68,70,72,73,75–79]
Role–Physical	.85	.74	.83–.98[65–68,70,72,73,75–79]
Bodily Pain	.95	.91	.79–.96[65–68,70,72,73,75–79]
General Health	.85	.80	.72–.95[65–68,70,72,73,75–79]
Vitality	.85	.91	.66–.96[65–68,70,72,73,75–79]
Social Functioning	.67	.84	.39–.98[65–68,70,72,73,75–79]
Role–Emotional	.91	.89	.78–.99[65–68,70,72,73,75–79]
Mental Health	.84	.93	.72–.95[65–68,70,72,73,75–79]
Unified Parkinson Disease Rating Scale			
Mentation, Behavior, and Mood	.64	.67	.79[80]
Activities of Daily Living	.75	.80	.85–.92[80–83]
Motor Examination	.87	.88	.88–.95[10,81–85]
Total score	.88	.90	.94–.96[10,86]

Results

Table 1 reports internal consistency for all multiple-item tests used in this study. All tests met the criterion of Cronbach alpha being .70 or greater, with the exception of day 1 for the Social Functioning subscale of the SF-36 and both days for the Mentation, Behavior, and Mood subscale of the UPDRS. Internal consistency from previous studies also is reported in Table 1. In previous studies, both the SF-36 Vitality and Social Functioning subscales have had internal consistency values less than .70.

Table 2 reports means, standard deviations, and confidence intervals from the first testing day, as well as the ICCs and MDC$_{95}$ values for all tests and measures administered in this study. The 6MWT was the only test that demonstrated statisti-

cally higher retest values (*t*=−2.15, *P*<.04), indicating that learning could be a factor for this test.

The BBS and ABC Scale were the most reliable of the balance measures, with MDC$_{95}$ values of 5 and 13, respectively. The BBS and ABC Scale both demonstrated a right-skewed distribution due to a ceiling effect on these scales.

Comfortable and fast gait speeds had the highest test-retest reliability, normal distributions, and MDC$_{95}$ values of 0.18 and 0.25 m/s. The 6MWT had excellent test-retest reliability and a normal distribution, but a large standard deviation that created a high MDC$_{95}$ value of 82 m. The TUG displayed a right-skewed distribution, but its test-retest reliability was low compared with the 6MWT and gait speeds.

All 8 subscales of the SF-36 had a test-retest reliability of .80 or above, except for the Social Functioning subscale. The Physical Functioning subscale, the scale most often used by physical therapists, had an MDC$_{95}$ value of 28% in our sample.

The UPDRS test-retest reliability values of .89 to .93 for the 3 subscales and total score were high, with MDC$_{95}$ values of 2, 4, 11, and 13, respectively. Minimal detectable change values for the UPDRS Mentation, Behavior, and Mood subscale should be used with caution due to its low internal consistency.

Discussion

The convenience sample of people with parkinsonism who participated in this study may be similar to patients with parkinsonism seen in outpatient clinics and wellness pro-

Table 2.
Sample Sizes, Means, Standard Deviations, 95% Confidence Intervals (CIs), Intraclass Correlation Coefficients (ICCs) for Test-Retest Reliability, and Minimal Detectable Changes (MDCs) for Balance and Ambulation Tests, a Quality of Life Measure, and a Disease Severity Rating Scale in People With Parkinsonism

Test Performed	n	Mean (SD)[a]	95% CI[a]	ICC[b]	MDC$_{95}$
Balance tests					
Berg Balance Scale (0–56 points)	37	50 (7)	47–52	.94	5
Activities-specific Balance Confidence Scale (%)	36	70 (19)	64–77	.94	13
Functional Reach Test (cm)					
Forward	37	21 (6)	18–23	.73	9
Backward	36	14 (5)	13–16	.67	7
Romberg Test (s)					
Eyes open	37	58 (10)	55–62	.86	10
Eyes closed	37	54 (17)	48–60	.84	19
Sharpened Romberg Test (s)					
Eyes open	37	39 (25)	30–47	.70	39
Eyes closed	37	15 (22)	8–23	.91	19
Ambulation tests					
Six-Minute Walk Test (m)	37	316 (142)	269–364	.96	82
Timed "Up & Go" Test (s)	37	15 (10)	12–19	.85	11
Gait speed (m/s)					
Comfortable	36	1.16 (.34)	1.04–1.27	.96	0.18
Fast	36	1.47 (.51)	1.30–1.64	.97	0.25
36-Item Short-Form Health Survey (0–100 points)					
Physical Functioning	36	57 (23)	49–65	.80	28
Role–Physical	36	47 (41)	33–61	.85	45
Bodily Pain	36	68 (27)	59–77	.89	25
General Health	36	59 (26)	50–67	.85	28
Vitality	36	52 (20)	45–59	.88	19
Social Functioning	36	83 (20)	76–90	.71	29
Role–Emotional	36	75 (40)	61–89	.84	45
Mental Health	36	76 (16)	70–81	.83	19
Unified Parkinson Disease Rating Scale (points)					
Mentation, Behavior, and Mood (0–16)	36	2 (2)	2–3	.89	2
Activities of Daily Living (0–52)	36	12 (6)	10–14	.93	4
Motor Examination (0–108)	37	19 (12)	15–23	.89	11
Total score (0–176)	36	33 (16)	28–38	.91	13

[a] Means, standard deviations, and 95% CIs are from first day of testing.
[b] ICC (3,1): Berg Balance Scale, Activities-specific Balance Confidence Scale, Romberg Test, Sharpened Romberg Test, Six-Minute Walk Test, 36-Item Short-Form Health Survey, and Unified Parkinson Disease Rating Scale. ICC (3,2): Functional Reach Test, Timed "Up & Go" Test, and gait speed.

grams but may have less severe PD than patients with the disease seen in long-term care and acute care inpatient facilities. Minimal detectable change values could vary not only by a disease but also by stage of the disease. The BBS, ABC Scale, SRT with eyes closed, 6MWT, and gait speed tests all demonstrated test-retest reliability values above .90. The MDCs calculated from these test-retest values are considered dependable. Functional tests with test-retest reliability values below .90 (forward and backward functional reach, RT with eyes open and eyes closed, SRT with eyes open, TUG) should be used with caution in people with parkinsonism. All of the SF-36 subscales except the Social Functioning and Role–Physical sub-

scales and all of the UPDRS subscales except the Mentation, Behavior, and Mood and ADL subscales had internal consistency and test-retest reliability values above .80, indicating that the scales measure one concept and that the MDCs are trustworthy.

Balance Testing

Internal consistency for the BBS in this study was similar to findings of a previous study of people with PD.[10] Our study's high test-retest reliability and calculated MDC_{95} values were similar to values reported in current literature (2–12) for people with various disabilities.[9,13-15,19] An MDC_{95} value of 5 on the BBS for people with parkinsonism is useful to physical therapists.

The test-retest reliability value of .73 for forward functional reach in this study was within the wide range of .42 to .93 reported in 2 previous studies of subjects with PD.[15,22] The calculated MDC_{95} of 9 cm for forward functional reach is between the values of 4 to 12 cm calculated from the previous literature for subjects with PD.[15,22] The low test-retest reliability value of .67 for backward functional reach, with a calculated MDC_{95} value of 7 cm, indicates that this test should be used with caution. Our test-retest reliability values for the SRT with eyes open and closed were slightly lower than the values obtained for the SRT with eyes open and closed in a previous study of elderly women who were healthy.[17] No previous research reports MDC values on these tests for individuals with parkinsonism. Many subjects reached the 60-second ceiling on the RT and SRT with eyes open. A floor effect was seen for the SRT with eyes closed, but this test had higher test-retest reliability values than the SRT with eyes open or the RT with eyes open and closed. Due to the low reliability of scores obtained for forward and backward functional reach and for the RT and

SRT (except for the SRT with eyes closed) in this study, these tests should be used cautiously as a measure of responsiveness to change in this population.

The ABC Scale had excellent internal consistency and test-retest reliability, with values being higher than those reported in the previous literature.[19,28-30] The MDC_{95} value of 13% in our study fell below the 18% to 38% calculated from the previous literature for other patient populations. A change score of 13% or greater should be used for people with parkinsonism.

Ambulation Testing

The test-retest reliability and MDC_{95} values obtained for 6MWT in this study fell within the range found in current literature.[31-41,43,44,47,48] None of these studies, however, assessed individuals with parkinsonism. The MDC_{95} value of 82 m was larger than desired due to a large standard deviation resulting from a wide range of disease severity of the participants on the H&Y. Even though the MDC_{95} value was high, test-retest reliability on the 6MWT for people with PD was excellent. Thus, an MDC_{95} value of 82 m is valid for clinicians using the 6MWT in individuals with parkinsonism. Future studies with greater numbers of patients in each H&Y stage will determine whether the standard deviation decreases secondary to better homogeneity of the group by stage. If so, separate MDCs on the 6MWT should be determined for each stage of the disease. Future researchers should check this functional test for learning effects. The effects of learning on the 6MWT found in this study, although significant, were small.

Test-retest reliability values obtained for the TUG in this study fell within the range of reliability values found in previous research studies.[15,50,52] The MDC_{95} values were higher than

desired but fell within the range of values reported in the current literature.[50,52] The mean score of 15 seconds on the TUG in this study would make a change score of 11 seconds or better unrealistic for the majority of the group. An MDC study based on each of the H&Y stages may decrease the standard deviation and subsequently provide lower MDCs on the TUG.

Test-retest reliability values for comfortable gait speed in this study fell within the values previously reported[13,21,32,37,53,55-58,60] and were higher than the reliability values obtained in the only other study reporting test-retest reliability for people with PD.[15] Our calculated MDC_{95} value of 0.18 m/s fell within the range of values reported in the literature[13,21,32,37,53,55-58,60] and was similar to the SDD value of .19 m/s reported in a previous study of people with PD.[15] The reliability and MDC_{95} values obtained for fast gait speed in the current study were similar to values reported in previous research.[21,32,37,57,58,60] The MDC_{95} values calculated for the gait speed tests in this study are valid for individuals with PD. Of the 4 ambulation tests presented, clinicians should consider using both the comfortable and fast gait speeds to measure responsiveness to change over time because of the high test-retest reliability, normal distribution, and useful MDC scores in people with PD.

SF-36

Internal consistency values for all 8 SF-36 subscales in this study fell within the Cronbach alpha values reported in previous research.[65-68,70,72,73,75-79] Similar to previous research, the Social Functioning scale had the poorest internal consistency (Tab. 1).

The test-retest reliability and MDC_{95} values (Tab. 1) calculated for all of the SF-36 subscales in this study fell

with the ranges reported in the previous literature.[63-74,76-79] None of these studies, however, assessed individuals with parkinsonism. The large MDC$_{95}$ values can be attributed to the broad diversity of populations tested. Each subscale of the SF-36 can be used independently. Therapy may improve a patient's quality of life as measured by the Physical Functioning subscale (10 items) and the Bodily Pain subscale (2 items), and these SF-36 subscales should be utilized by therapists. A change of 28/100 or higher on the Physical Functioning subscale and a change of 25/100 or higher on the Bodily Pain subscale would demonstrate an improvement in these quality-of-life dimensions.

UPDRS

Internal consistency for the Motor Examination subscale of the UPDRS in this study was similar to the ranges reported in previous studies.[10,81-85] Cronbach alpha values for the UPDRS subscales and total score in this study all fell slightly below reported values in the literature, which may be due to the large sample sizes used in previous studies.[10,82,83,86] Internal consistency of the Mentation, Behavior, and Mood subscale of the UPDRS was below the acceptable level of .70, and this subscale should be used with caution in measuring change over time, despite acceptable test-retest scores.

The test-retest reliability and MDC$_{95}$ values for the UPDRS subscales and total score were similar to values obtained in previous studies.[15,87,88] The MDC$_{95}$ values of 2 for the Mentation, Behavior, and Mood subscale and 4 for the ADL subscale in this study were the same as the values obtained for those subscales in a previous study that examined 400 patients with early stage, mild PD, but 4 points higher for the Motor Examination subscale and total score.[87] The higher MDC$_{95}$ values in this study may have been due to the smaller sample size and wider representation of PD severity.

Conclusion

Therapists have evaluation choices for measuring balance, ambulation, quality of life, and disease severity when assessing change over time in patients with chronic disease. The MDCs found for the BBS, ABC Scale, SRT with eyes closed, comfortable and fast gait speeds, 6MWT, SF-36 subscales (except Social Functioning), and UPDRS ADL and Motor Examination subscales and total score will be useful to therapists working with patients with parkinsonism in rehabilitation and wellness programs to determine whether change is due to testing error or is a result of intervention techniques. These values also help therapists interpret literature comparing statistical significance with meaningful clinical change. Test-retest reliability studies with larger samples by stage of PD and for patients with Parkinson-plus syndromes will help further define MDC values.

References

1 Iyer LV, Haley SM, Watkins MP, Dumas HM. Establishing minimal clinically important differences for scores on the Pediatric Evaluation of Disability Inventory for inpatient rehabilitation. *Phys Ther.* 2003;83: 888–897.

2 Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Phys Ther.* 1996;76:359–367.

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

4 Portney L, Watkins M. *Foundations of Clinical Research: Applications to Practice.* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall Inc; 2000.

5 Teixeira da Cunha-Filho I, Henson H, Qureshy H, et al. Differential responses to measures of gait performance among healthy and neurologically impaired individuals. *Arch Phys Med Rehabil.* 2003; 84:1774–1779.

6 Palombaro KM, Craik RL, Mangione KK, Tomlison JD. Determining meaningful changes in gait speed after hip fracture. *Phys Ther.* 2006;86:809–816.

7 Berg KO, Wood-Dauphinée S, Williams JI, Gayton D. Measuring balance in the elderly: preliminary development of an instrument. *Physiother Can.* 1989;41: 304–311.

8 Berg KO, Maki BE, Williams JI, et al. Clinical and laboratory measures of postural balance in an elderly population. *Arch Phys Med Rehabil.* 1992;73:1073–1080.

9 Berg KO, Wood-Dauphinée S, Williams JI. The Balance Scale: reliability assessment with elderly residents and patients with an acute stroke. *Scand J Rehabil Med.* 1995; 27:27–36.

10 Brusse KJ, Zimdars S, Zalewski KR, Steffen TM. Testing functional performance in people with Parkinson disease. *Phys Ther.* 2005;85:134–140.

11 Halsaa K, Brovold T, Graver V, et al. Assessments of interrater reliability and internal consistency of the Norwegian version of the Berg Balance Scale. *Arch Phys Med Rehabil.* 2007;88:94–98.

12 Mao HF, Hsueh IP, Tang PF, et al. Analysis and comparison of the psychometric properties of three balance measures for stroke patients. *Stroke.* 2002;33:1022–1027.

13 Liston R, Bouwer B. Reliability and validity of measures obtained from stroke patients using the Balance Master. *Arch Phys Med Rehabil.* 1996;77:425–430.

14 Newstead A, Hinman M, Tomberlin J. Reliability of the Berg Balance Scale and Balance Master limits of stability tests for individuals with brain injury. *Journal of Neurologic Physical Therapy.* 2005;29(1): 18–23.

15 Lim L, van Wegen E, de Goede C, et al. Measuring gait and gait-related activities in Parkinson's patients' own home environment: a reliability, responsiveness, and feasibility study. *Parkinsonism Relat Disord.* 2005;11:19–24.

16 Duncan P, Weiner D, Chandler J, Studenski S. Functional Reach: a new clinical measure of balance. *J Gerontol Med Sci.* 1990;45:M192–M197.

17 Franchignoni F, Tesio L, Martino M, Ricupero C. Reliability of four simple, quantitative tests of balance and mobility in healthy elderly females. *Aging Clin Exp Res.* 1998;10:26–31.

18 Hageman P. Gait characteristics of healthy elderly: a literature review. *Issues on Aging.* 1995;18:14–18.

19 Holbein-Jenny M, Billek-Sawhney B, Beckman E, Smith T. Balance in personal care home residents: a comparison of the Berg Balance Scale, the Multi-Directional Reach Test, and the Activities-specific Balance Confidence Scale. *Journal of Geriatric Physical Therapy.* 2005;28(2):48–53.

20 Schenkman M, Cutson TM, Kuchibhatla M, et al. Reliability of impairment and physical performance measures for persons with Parkinson's disease. *Phys Ther.* 1997;77:19–26.

21 Sherrington C, Lord S. Reliability of simple portable tests of physical performance in older people after hip fracture. *Clin Rehabil.* 2005;19:496–504.

22 Smithson F, Morris ME, Iansek R. Performance on clinical tests of balance in Parkinson's disease. *Phys Ther.* 1998;78:577–592.

23 Weiner D, Duncan P, Chandler J, Studenski S. Functional Reach: a marker of physical frailty. *J Am Geriatr Soc.* 1992;40:203–207.

24 Marsh AP, Rejeski JW, Hutton SL, et al. Development of a lateral mobility task to identify individuals at risk for mobility disability and functional decline. *J Aging Phys Act.* 2005;13:363–381.

25 Kammerlind A, Larsson P, Ledin T, Skargren E. Reliability of clinical balance tests and subjective ratings in dizziness and disequilibrium. *Advances in Physiotherapy.* 2005;7:96–107.

26 Hamilton K, Kantor L, Magee L. Limitations of postural equilibrium tests for examining simulator sickness. *Aviat Space Environ Med.* 1989;60:246–251.

27 Black F, Wall C, Rockette H, Kitch R. Normal subject postural sway during the Romberg Test. *Am J Otolaryngol.* 1982;3:309–318.

28 Parry S, Steen N, Galloway S, et al. Falls and confidence-related quality of life outcome measures in an older British cohort. *Postgrad Med J.* 2001;77:103–108.

29 Miller WD, AB, Speechley M. Psychometric properties of the Activities-specific Balance Confidence Scale among individuals with a lower-limb amputation. *Arch Phys Med Rehabil.* 2003;84:656–661.

30 Powell L, Myers A. The Activities-specific Balance Confidence (ABC) Scale. *J Gerontol A Biol Sci Med Sci.* 1995;50:M28–M34.

31 Eng J, Dawson A, Chu K. Submaximal exercise in persons with stroke: test-retest reliability and concurrent validity with maximal oxygen consumption. *Arch Phys Med Rehabil.* 2004;85:113–118.

32 Flansbjer U, Holmback A, Downham D, et al. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med.* 2005;37:75–82.

33 Mossberg K. Reliability of a timed walk test in persons with acquired brain injury. *Am J Phys Med Rehabil.* 2003;82:385–392.

34 Andersson C, Asztalos L, Mattsson E. Six-minute walk test in adults with cerebral palsy. a study of reliability. *Clin Rehabil.* 2006;20:488–495.

35 Gylfadottir S, Dallimore M, Dean E. The relation between walking capacity and clinical correlates in survivors of chronic spinal poliomyelitis. *Arch Phys Med Rehabil.* 2006;87:944–952.

36 Fry D, Pfalzer L. Reliability of four functional tests and rating of perceived exertion in persons with multiple sclerosis. *Physiother Can.* 2006;58:212–219.

37 Van Loo M, Moseley A, Bosman J, et al. Test-retest reliability of walking speed, step length, and step width measurements after traumatic brain injury: a pilot study. *Brain Inj.* 2004;18:1041–1048.

38 Hamilton D, Haennel R. Validity and reliability of the 6-Minute Walk Test in a cardiac rehabilitation population. *J Cardiopulmonary Rehabil.* 2000;20:156–164.

39 Guyatt GH, Thompson PJ, Berman LB, et al. How should we measure function in patients with chronic heart and lung disease? *J Chronic Dis.* 1985;38:517–524.

40 Demers C, McKelvie R, Negassa A, Yusuf S. Reliability, validity, and responsiveness of the Six-Minute Walk Test in patients with heart failure. *Am Heart J.* 2001;142:698–703.

41 Montgomery PS, Gardner AW. The clinical utility of a six-minute walk test in peripheral arterial occlusive disease patients. *J Am Geriatr Soc.* 1998;46:706–711.

42 Opasich C, Pinna G, Mazza A, et al. Reproducibility of the Six-Minute Walking Test in patients with chronic congestive heart failure: practical implications. *Am J Cardiol.* 1998;81:1497–1500.

43 Brooks D, Solway S, Weinacht K, et al. Comparison between an indoor and an outdoor 6-minute walk test among individuals with chronic obstructive pulmonary disease. *Arch Phys Med Rehabil.* 2003;84:873–876.

44 Sciurba F, Criner G, Lee S, et al. Six-minute walk distance in chronic obstructive pulmonary disease. *Amer J Respir Crit Care Med.* 2003;167:1522–1527.

45 Roomi J, Johnson M, Waters K, et al. Respiratory rehabilitation, exercise capacity, and quality of life in chronic airways disease in old age. *Age Ageing.* 1996;25:12–17.

46 Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the Six Minute Walk Test in chronic lung disease patients. *Amer J Respir Crit Care Med.* 1997;155:1278–1282.

47 Harada ND, Chiu V, Stewart AL. Mobility-related function in older adults: assessment with a 6-minute walk test. *Arch Phys Med Rehabil.* 1999;80:837–841.

48 King MB, Judge JO, Whipple R, Wolfson L. Reliability and responsiveness of two physical performance measures examined in the context of a functional training intervention. *Phys Ther.* 2000;80:8–16.

49 Noren A, Bogren U, Bolin J, Stenstrom C. Balance assessment in patients with peripheral arthritis: applicability and reliability of some clinical assessments. *Physiother Res Int.* 2001;6:193–204.

50 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.

51 Podsiadlo D, Richardson S. The Timed "Up and Go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc.* 1991;39:142–148.

52 Campbell C, Rowse J, Ciol M, Shumway Cook A. The effect of cognitive demand on timed up and go performance in older adults with and without Parkinson disease. *Neurol Rep.* 2003;27:2–6.

53 Green J, Forster A, Young J. Reliability of gait speed measured by a timed walking test in patients one year after stroke. *Clin Rehabil.* 2002;16:306–314.

54 Dawes H, Smith C, Collett J, et al. A pilot study to investigate explosive leg extensor power and walking performance after stroke. *J Sports Sci Med.* 2005;4:556–562.

55 Cunha I, Lim P, Henson H, et al. Performance-based gait tests for acute stroke patients. *Am J Phys Med Rehabil.* 2002;81:848–856.

56 Jette AM, Jette DU, Ng J, et al; for the Musculoskeletal Impairment (MIS) Study Group. Are performance-based measures sufficiently reliable for use in multicenter trials? *J Gerontol A Biol Sci Med Sci.* 1999;54:M3–M6.

57 Boonstra A, Fidler V, Eisma W. Walking speed of normal subjects and amputees: aspects of validity of gait analysis. *Prosthet Orthot Int.* 1993;17:78–82.

58 Lusardi M, Pellechia G, Schulman M. Functional performance in community living older adults. *Journal of Geriatric Physical Therapy.* 2003;26:14–22.

59 Rolland Y, Cesari M, Miller M, et al. Reliability of the 400-m Usual-Pace Walk Test as an assessment of mobility limitation in older adults. *J Am Geriatr Soc.* 2004;52:972–976.

60 Fransen M, Crosbie J, Edmonds J. Reliability of gait measurements in people with osteoarthritis of the knee. *Phys Ther.* 1997;77:944–953.

61 Kennedy DM, Stratford PW, Wessel J, et al. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskelet Disord.* 2005;6:3.

62 Sherrington C, Lord S. Home exercise to improve strength and walking velocity after hip fracture: a randomized controlled trial. *Arch Phys Med Rehabil.* 1997; 78:208–212.

63 Enloe LJ, Shields RK. Evaluation of health-related quality of life in individuals with vestibular disease using disease-specific and general outcome measures. *Phys Ther.* 1997;77:890–903.

64 Suris A, Borman PD, Lind L, Kashner TM. Aggression, impulsivity, and health functioning in a veteran population: equivalency and test-retest reliability of computerized and paper-and-pencil administrations. *Comput Human Behav.* 2007;23:97–110.

65 Dougherty C, Dewhurst T, Nichol W, Spertus J. Comparison of three quality of life instruments in stable angina pectoris: Seattle Angina Questionnaire, Short Form Health Survey (SF-36), and Quality of Life Index–Cardiac Version III. *J Clin Epidemiol.* 1998;51:569–575.

66 Khoudri I, Ali Zeggwagh A, Abidi K, et al. Measurement properties of the SF-36 and health-related quality of life measures after intensive care in Morocco. *Acta Anaesthesiol Scand.* 2007;51:189–197.

67 Lin MR, Hwang HF, Chen CY, Chiu WT. Comparisons of the brief form of the World Health Organization Quality of Life and Short Form-36 for persons with spinal cord injuries. *Am J Phys Med Rehabil.* 2007;86:104–113.

68 Dorman P, Slattery J, Farrell B, et al. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. *Stroke.* 1998;29:63–68.

69 Ruta D, Hurst N, Kind P, et al. Measuring health status in British patients with rheumatoid arthritis: reliability, validity, and responsiveness of the Short Form 36-Item Health Survey (SF-36). *Br J Rheumatol.* 1998;37:425–436.

70 Thumboo J, Feng P, Boey M, et al. Validation of the Chinese SF-36 for quality of life assessment in patients with systemic lupus erythematosus. *Lupus.* 2000;9:708–712.

71 Marx R, Menezes A, Horovitz L, et al. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol.* 2003;56:730–735.

72 Andresen E, Bowley N, Rothenberg B, et al. Test-retest performance of a mailed version of the Medical Outcomes Study 36-Item Short-Form Health Survey among older adults. *Med Care.* 1996;34:1165–1170.

73 Ruta D, Abdalla M, Garratt A, et al. SF 36 health survey questionnaire, I: reliability in two patient based studies. *Qual Health Care.* 1994;3:180–185.

74 Campbell K, Rohlman D, Storzbach D, et al. Test-retest reliability of psychological and neurobehavioral tests self-administered by computer. *Assessment.* 1999;6:21–32.

75 Li L, Wang H, Shen Y. Chinese SF-36 health survey: translation, cultural adaptation, validation, and normalization. *J Epidemiol Community Health.* 2003;57:259–263.

76 Gonzalez N, Quintana J, Arostegui I, et al. Translation and psychometric testing of the Basque version of the SF-36 health survey. *Qual Life Res.* 2005;14:549–554.

77 Sanson-Fisher R, Perkins J. Adaptation and validation of the SF-36 health survey for use in Australia. *J Clin Epidemiol.* 1998; 51:961–967.

78 Fukuhara S, Bito S, Green J, et al. Translation, adaptation, and validation of the SF-36 health survey for use in Japan. *J Clin Epidemiol.* 1998;51:1037–1044.

79 Brazier J, Harper R, Jones N, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *Br Med J.* 1992;305:160–164.

80 Martinez-Martin P, Forjaz M. Metric attributes of the Unified Parkinson's Disease Rating Scale 3.0 battery, part I: feasibility, scaling assumptions, reliability, and precision. *Mov Disord.* 2006;21:1182–1188.

81 Martignoni E, Franchignoni F, Pasetti C, et al. Psychometric properties of the Unified Parkinson's Disease Rating Scale and of the Short Parkinson's Evaluation Scale. *Neuro Sci.* 2003;24:190–191.

82 Martinez-Martin P, Prieto L, Forjaz M. Longitudinal metric properties of disability rating scales for Parkinson's disease. *Value in Health.* 2006;9:386–393.

83 Van Hilten J, van der Zwan A, Zwinderman A, Roos R. Rating impairment and disability in Parkinson's disease: evaluation of the Unified Parkinson's Disease Rating Scale. *Mov Disord.* 1994;9:84–88.

84 Stebbins G, Goetz C. Factor structure of the Unified Parkinson's Disease Rating Scale: motor examination section. *Mov Disord.* 1998;13:633–636.

85 Stebbins G, Goetz C, Lang A, Cubo E. Factor analysis of the motor section of the unified Parkinson's Disease Rating Scale during the off-state. *Mov Disord.* 1999;14: 585–589.

86 Martinez-Martin P, Gil-Nagel A, Garcia M. Unified Parkinson's Disease Rating Scale characteristics and structure. *Mov Disord.* 1994;9:76–83.

87 Siderowf A, McDermott M, Kieburtz K, et al. Test-retest reliability of the Unified Parkinson's Disease Rating Scale in patients with early Parkinson's disease: results from a multicenter clinical trial. *Mov Disord.* 2002;17:758–763.

88 Metman L, Myre B, Verwey N, et al. Test-retest reliability of UPDRS-III, dyskinesia scales, and timed motor tests in patients with advanced Parkinson's disease: an argument against multiple baseline assessments. *Mov Disord.* 2004;19:1079–1084.

89 Billek-Sawhney B, Gay J. The Functional Reach Test: Are 3 trials necessary? *Topics in Geriatric Rehabilitation.* 2005;21: 144–148.

90 Society AT. Guidelines for the Six-Minute Walk Test. *Amer J Resp Crit Care Med.* 2002;166:111–117.

91 Ware J. *SF-36 Health Survey: Manual and Interpretation Guide.* Boston, Mass: Nimrod Press; 1993.

92 Goetz C, Stebbins G, Chmura T, et al. Teaching tape for the motor section of the Unified Parkinson's Disease Rating Scale. *Mov Disord.* 1995;10:263–266.

93 Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. *Neurology.* 1967;17:427–442.