Research Report

Reliability of Impairment and Physical Performance Measures for Persons With Parkinson's Disease

Background and Purpose. Parkinson's disease (PD) is characterized by rigidity, postural instability, bradykinesia, and tremor, as well as other musculoskeletal impairments and functional limitations. The purpose of this investigation was to determine the reliability and stability of measures of impairments and physical performance for people in the early and middle stages of PD. **Subjects.** Thirteen men and 2 women in Hoehn and Yahr stages 2 and 3 of PD participated. Their mean age was 74.5 years (SD=5.7, range=64–84). **Methods.** Thirteen impairment-level variables and 8 physical performance variables were measured. Measurements were taken on two consecutive days and again a week later on the corresponding two consecutive days. Reliability and stability were assessed using analysis of variance and intraclass correlation coefficients (ICCs). **Results.** Test-retest reliability (ICCs) of variables ranged from .69 (hamstring muscle length) to .97 (lumbar flexion). Intraclass correlation coefficients were .85 or greater for 10 of the variables. **Conclusions and Discussion.** The results suggest that in the early and middle stages of PD, many of the measures of impairment and physical performance are relatively stable. [Schenkman M, Cutson T, Kuchibhatla M, et al. Reliability of impairment and physical performance measures for persons with Parkinson's disease. *Phys Ther.* 1997;77:19–27.]

Key Words: *Neuromusculoskeletal disorders, Parkinson's disease; Tests and measurements, functional.*

Margaret Schenkman

Toni M Cutson

Maggie Kuchibhatla

Julie Chandler

Carl Pieper

Parkinson's disease (PD) is a disease of older individuals affecting 100 to 150 of every 100,000 people in the United States, with a prevalence of about 1% of those over the age of 60 years.[1] Parkinson's disease results from neurotransmitter imbalances associated with degeneration of the substantia nigra.[2,3] The primary impairments typically are rigidity, bradykinesia, tremor, and postural instability.[2,3] In addition, people with PD often have stooped, flexed posture, characterized by excessive thoracic kyphosis and loss of lumbar lordosis.[4] Mobility of the neck, torso, and extremities is also lost. Functional limitations in bed mobility, transfers, and gait may become severely disabling as the disease progresses.

Physical therapy for persons in the early stages of PD is directed at correcting musculoskeletal impairments and improving physical performance.[4,5] Reliable and valid measures of the impairments and functional limitations associated with PD are important for clinical practice and are a prerequisite to clinical investigations of PD.

Various approaches have been reported for rating the symptoms of people with PD.[6–10] Recently, a few investigators have reported on the reliability or validity of scales used to rate PD[10,11] or have compared the performance of different rating scales.[12,13] It is evident from such studies that the scales often measure different aspects of PD and that it is not feasible to compare patients who have been rated by different scales.[12,13]

Most measures used to quantify impairments or physical performance of persons with PD are global measures that characterize the overall effects of the disease (eg, the Hoehn and Yahr[14] score used to stage the patient), rely heavily on the patient's self-report (eg, Unified Parkinson's Disease Rating Scale,[8] Northwestern Rating Scale[9]), or emphasize the direct effects of the disease, including tremor and rigidity.[15] Most of the scales also are constructed to be used across the full range of stages of PD. None of the available measures quantify the patient's musculoskeletal impairments or provide performance-based measures of function across a range of activities associated with activities of daily living. Reliable measurements obtained by physical examination may be particularly difficult to obtain for persons with PD because of the fluctuating nature of the disease. Signs and symptoms vary with the time of day, medication schedule, and anxiety level.[5]

This study was designed to examine the reliability of performance-based measures of particular relevance to the functional mobility of persons in the early and middle stages of PD in order to establish the utility of these measures in research and clinical investigations. The measures investigated were chosen to represent a range of impairments (eg, force production, range of motion [ROM], spinal configuration [lumbar lordosis, thoracic kyphosis]) and of physical performance (eg, balance control, transfers, walking).

The purposes of the investigation were (1) to determine whether there were systematic variations in the variables by day or week of testing and (2) to determine the test-retest reliability of measurements obtained for the variables investigated.

M Schenkman, PhD, PT, is Associate Professor, Graduate Program in Physical Therapy, Senior Fellow, Center for the Study of Aging and Human Development, and Co-Director, Claude D Pepper Older Americans Independence Center, Duke University, Durham, NC 27710. Address all correspondence to Dr Schenkman at the Center for the Study of Aging and Human Development, Duke University Medical Center, PO Box 3003, Durham, NC 27710 (USA) (ms@geri.duke.edu).

TM Cutson, MD, is Associate Professor, Department of Family Practice, The University of Michigan, Ann Arbor, MI 48109.

M Kuchibhatla, PhD, is Statistician, Claude D Pepper Older Americans Independence Center, Center for the Study of Aging and Human Development, Duke University Medical Center.

J Chandler, PhD, PT, is Senior Epidemiologist, Merck Research Laboratories, Blue Bell, PA 19486.

C Pieper, DPH, is Director, Analysis Core, Claude D Pepper Older Americans Independence Center, Center for the Study of Aging and Human Development, and Assistant Professor, Department of Community and Family Medicine, Division of Biometry, Duke University Medical Center.

This study was approved by the Institutional Review Board of Duke University Medical Center.

This work was supported by the National Institutes of Health, National Institute on Aging, Claude D Pepper Older Americans Independence Center Grant No. 5 P60 AG 11268 and by the National Institutes of Health, National Center for Research Resources, General Clinical Research Centers Program Grant No. MO1-RR-30.

This article was submitted December 20, 1995, and was accepted August 28, 1996.

Method

Raters

Two physical therapists with 7 and 2 years of clinical experience, respectively, and two research assistants participated in this study. They were part of a measurement team that provided all measures for four intervention studies of the Claude D Pepper Older Americans Independence Center (OAIC) at the Duke University Center for the Study of Aging and Human Development. The measurement team performed these measures on a routine basis and had measured more than 100 subjects at the time that this study was conducted. Prior to rating subjects for the OAIC Measurement Core, all raters underwent a training session, including performance of measures and consistent use of the protocols, and a checkout with a physical therapist who was experienced in the use of these measures and who was a coinvestigator in one of the OAIC studies. The same rater evaluated the same participant for each of the four sessions whenever possible. Because this reliability study was part of a large, ongoing study, this was not always feasible.

Subjects

Subjects were included who had been diagnosed by a neurologist as having PD, who were able to ambulate independently, and who had been on a stable drug regimen for at least 1 month. Subjects were excluded if their Folstein Mini Mental State Examination[16] score was less than 23, if they had been hospitalized within the previous 3 months, or if they had symptoms of another neurological disease (eg, cerebrovascular accident). The subjects were recruited from Durham, NC, and the surrounding areas. The subjects who participated in this investigation were a subset of the participants in a larger randomized clinical trial examining the effects of exercise in persons with PD. All subjects signed an informed consent statement prior to participation.

Thirteen men and 2 women who were independent ambulators participated. Characteristics of the participants are shown in Table 1. The mean age of the subjects was 74.5 years (SD=5.7, range=64–84). Eight participants were in stage 2 of PD, according to Hoehn and Yahr's staging for PD,[14] and 7 participants were in Hoehn and Yahr stage 3. By definition, patients in Hoehn and Yahr stage 2 are independent ambulators and have unilateral symptoms and intact balance; patients in stage 3 are independent ambulators and have more severe bilateral signs of the disease and impaired balance. Two of the subjects used assistive devices (cane or walker) some of the time. The other subjects walked without an assistive device. Income and educational level for this group were high. Four of the subjects reported having annual incomes of $20,000 to $50,000; the remaining subjects reported having incomes of greater than $50,000. Their mean years of education was 16.8 (SD=2.4, range=12–21).

Subjects underwent a brief medical examination prior to entering the study. At that time, medication history, the status of tremors, rigidity, and Hoehn and Yahr staging of disease were determined. Subjects were tested on two consecutive days and again a week later on the corresponding two consecutive days.

Variables

The variables measured in this study included ROM (9 variables), spinal configuration (2 variables), muscle force (2 variables), and physical performance (8 variables). Whenever possible, published protocols with established reliability were used. In some instances, when no protocol or reliability information was available, we report the findings of a preliminary reliability study of a cohort of community-dwelling elderly people who were without specific diseases.[17] A few of the measures were included for which no preliminary reliability study had been carried out and for which the current report represents the first reported reliability.

Procedure

All impairment-level variables (eg, ROM, muscle force) and one physical performance variable (the 6-minute walk) were measured by a physical therapist. The same therapist took these measurements for the four test sessions for 12 participants and for 75% of the test sessions for the remaining 3 subjects. All physical performance variables (except the 6-minute walk) were measured by a research assistant. The same research assistant took the measurements for four test sessions for 9 participants and for 62% of the test sessions for the remaining 6 subjects. The total battery of measures took between 45 minutes and 1½ hours to perform, depending on the participant's functional ability.

To minimize the effects of fluctuations of drug levels, subjects were requested to take PD medications at the same time on each day of testing, and the time of the test session was held constant. A research assistant called the subjects prior to each test session to remind them when they previously had taken their medications. During each test session, if the variables were measured more than once, the mean value was the variable used in the analysis. Because these measures were part of a battery of tests used in an intervention study, decisions were made to reduce respondent burden. Two trials were performed for primary outcomes, and the data were averaged. A single trial was performed for other variables.

Impairment measures. Measures related to the spine included spinal configuration, lumbar ROM, and cervi-

Table 1.
Characteristics of the Sample

Subject No.	Gender/Age (y)	No. of Years Since Diagnosis of Parkinson's Disease	Hoehn and Yahn[14] Score	Antiparkinsonian Medications	Folstein Mini Mental State Examination[18] Score	Use of Assistive Device	Tremor[a]	Rigidity[b]
1	M/80	6	3.0	Sinemet®	24	No	++	++
2	M/69	1	2.5	Sinemet®	27	No	No	+
3	F/75	11	2.5	Sinemet® Amantadine	30	Cane	+	No
4	M/84	8	2.5	Sinemet® Eldepryl®	26	No	No	++
5	F/80	20	2.5	Eldepryl®	29	No	++	++
6	M/69	4	2.0	Sinemet® Eldepryl®	30	No	+++	No
7	M/74	2	3.0	Sinemet®	29	No	++	++
8	M/64	1	3.0	Sinemet® Eldepryl®	28	No	No	++
9	M/79	5	3.0	Sinemet®	30	No	+	+
10	M/69	18	3.0	Sinemet® Eldepryl® Parlodel®	30	No	+	++
11	M/72	3	2.5	Sinemet® Eldepryl®	28	No	No	++
12	M/78	2	2.5	Sinemet® Eldepryl®	30	No	+++	++
13	M/79	7	3.0	Sinemet® Eldepryl®	28	Cane Walker	No	+
14	M/79	2	2.5	Sinemet® Eldepryl®	29	Cane	No	+
15	M/69	3	3.0	Sinemet® Eldepryl®	28	No	++	++

[a] +=mild; ++=moderate: present and visible, only occasionally bothersome; +++=severe: constant and very obvious.
[b] +=mild: only apparent with reinforcement; ++=moderate: present but joint can be fully extended with passive range of motion; +++=severe: limits available passive range of motion.

cal ROM. Extremity measures included ROM and muscle force.

The Debrunner kyphometer™* was used to measure spinal configuration (ie, thoracic kyphosis and lumbar lordosis).[18] Midpoints between T2–3, T11–12, and S1–2 were identified by palpation and marked. Subjects stood in their normal posture, looking straight ahead. Their feet were hip-width apart, and their arms rested by their sides. The blocks of the kyphometer spanned T2–3 and T11–12 for thoracic kyphosis and T11–12 and S1–2 for lumbar lordosis. A single trial was recorded for each variable. Excellent test-retest reliability for these measures has been reported for well-elderly subjects (intraclass correlation coefficients [ICCs]=.90 or above).[18] Interrater reliability was established in our laboratory[17] for a sample of well-elderly subjects prior to this investigation (ICCs=.95 for thoracic configuration and .96 for lumbar configuration).

The Back Range of Motion (BROM™) instrument† was used to determine lumbar ROM in a standing position. The spinous processes of T-12 and S-1 were palpated and marked. Resting position was recorded. A practice trial was conducted, and then two test trials were recorded and averaged for each variable. For flexion and extension, the upper contact point of the base of the BROM™ instrument was placed at S-1 and the sliding arm was placed at T-12. The degrees of maximum flexion and extension were read directly from the outer scale of the unit. For these measures, subjects were instructed to bend as far forward and backward as possible. For side bending, the positioning frame was placed at T-12 and

* Protek AG, Bern, Switzerland.

† Performance Attainment Associates, 958 Lydia Dr, Roseville, MN 55113.

subjects were instructed to side bend as far as possible. The degrees of maximum right and left side bending were read from the inclinometer. Test-retest reliability has been reported for all measures for a sample of elderly subjects with osteoarthritis (ICCs=.72–.94).[19]

The mobility of the cervical spine was measured with the subjects in a sitting position using the Cervical Range of Motion (CROM™) instrument.[†][20,21] With the CROM™ instrument on the head and the magnetic yoke on the chest, each subject's resting position was recorded. The subject was then instructed to move as far as possible into each plane of motion (flexion, extension, side bending, and rotation). The maximum degrees of motion for each trial were read from the appropriate inclinometer. Data from two trials were averaged for each motion. Test-retest and interrater reliability have been established for measurements of all planes of motion for asymptomatic individuals (ICCs=.76 or above).[20,21]

Standard goniometric techniques[22] were used to measure extremity ROM (ie, shoulder flexion and ankle dorsiflexion). Data were recorded from a single trial. Hamstring muscle length was determined with the subject positioned supine. The hip was flexed to 90 degrees, and then the knee was extended to the point of resistance. The angle between the tibia and the femur was recorded as a measure of hamstring muscle length.[23]

Extremity muscle force was measured using a hand-held dynamometer and a modification of published methods.[24,25] A Chatillon hand-held dynamometer[‡] was used to obtain isometric measurements of peak force. Ankle dorsiflexion force was measured with the subject in a sitting position (hip in about 90° of flexion, knee in 45° of flexion, and ankle in 30° of plantar flexion). The dynamometer was applied perpendicular to the metatarsal heads. Hip abduction was measured with the subject positioned supine and with the leg in a neutral position. The thigh-segment length from the greater trochanter to the position of the dynamometer was measured and used to calculate the actual force produced. A practice trial was conducted, followed by two test trials. The data were averaged. Prior to initiating this study, interrater reliability was established for asymptomatic elderly subjects for ankle dorsiflexion (ICC=.93, right side; ICC=.95, left side) and hip abduction (ICC=.89, right and left sides).[17]

Physical performance measures. Measures of physical performance included a series of tasks that required mobility of the spine, such as twisting, looking behind (functional axial rotation [FAR]), and walking.

Functional axial rotation[26] is a measure of combined spinal motions that we believe relate to functional abilities. The validity of the FAR measure is indicated by its correlation with physical performance measures that incorporate motion of the neck and back (canonical correlation coefficient $r=.60$, $P=.005$).[27] Functional axial rotation was assessed with the subject seated and the pelvis stabilized by Velcro® straps.[§] A hoop with symbols (numbers and letters) in 5-degree increments was suspended at eye level by two tripods, one in front of the subject and the other behind the subject. The headpiece of the CROM™ instrument was placed on the subject's head. The forward head arm of the unit was used as a pointer oriented toward the hoop. The subject was instructed to turn as far as possible to the right and then to the left and to report the farthest symbol that could be seen. The symbol with which the pointer aligned was recorded as FAR. A practice trial and two test trials were conducted, and the data were averaged. Excellent interrater and test-retest reliability have been reported (ICCs=.90 or above).[26]

Functional reach,[28] a measure of balance control, was measured with the subject in a standing position. A yardstick was positioned on the wall at the height of the acromion. The subject's dominant arm was held in 90 degrees of shoulder flexion. The subject was instructed to reach as far forward as possible without taking a step, and the distance reached was recorded. Two practice trials and three test trials were performed. Data from the three test trials were averaged. Excellent interrater reliability has been reported for a sample of elderly individuals (ICCs=.90 or above).[38]

A digital stopwatch was used to time the subject during movement from a supine position on a low treatment table to a standing position. The subject was instructed to move at his or her normal pace. The subject was then given the same instructions and requested to return to a supine position. One practice trial and two test trials were conducted for each variable, and the data were averaged.

The 360-degree turn measures the number of steps and the time required to turn around in place in a standing position. A digital stopwatch was used to time the task, and the number of steps was recorded. Each subject completed two test trials to the right side, and the data were averaged. Similarly, each subject completed two trials to the left side, and the data were averaged. Time (in seconds) and number of steps were recorded.

For the 6-minute walk, the subject was requested to walk at a comfortable pace for 6 minutes.[29] Distance was recorded for a single trial.

[‡] John Chatillon & Sons Inc, PO Box 35668, Greensboro, NC 27425-5668.

[§] Velcro USA Inc, 406 Brown Ave, PO Box 5218, Manchester, NH 03108.

Table 2.
Analysis-of-Variance Table for Parkinson's Disease Reliability Experiment[a]

Source	*df*	SS	MS	EMS
Between subjects	N−1 (15−1)	$N = 16\ k(\bar{X}_i - \bar{X})^2$ $\sum_{i=1}$	BMS	$\sigma_e^2 + k\,\sigma_T^2$
Within subjects	Nk−N (15×k−15)	$N = 16\ ks_i^2$ $\sum_{i=1}$	WMS	σ_e^2
Total	Nk−1	$(Nk-1)\ S^2$		

[a] k=total number of measurements per subject, N=number of subjects, $\bar{X}_i$=mean of the k readings for subject i, $\bar{X}$=mean of all readings, s_i^2=variance of the k readings for subject i, S^2=variance of all readings across all subjects, σ_e^2=expected within-person (error) variance, σ_T^2=expected between-person (true) variance, BMS=between mean squares, WMS=within mean squares, EMS=expected mean squares.

For the 10-m walk, a 10-m distance was marked on the floor.[30] The subject was instructed to walk at a comfortable pace. The subject began this test 5 m before the starting line and completed the test about 5 m after the finish line. Time was recorded from the time when the subject crossed the starting line to the time when he or she crossed the finish line. Each subject completed two test trials, and the data were averaged.

Data Analysis

Variation by day or week of testing. Each variable was measured during four different test sessions (days 1, 2, 8, and 9). We used a repeated-measures analysis of variance (ANOVA), with the repeated observations nested within subjects, to determine whether there were systematic fluctuations by day or week of testing. For each variable, data were included in the ANOVA and the reliability determinations only if data existed for all four test sessions (ie, listwise deletion was used in the analysis). This analysis allowed us to determine whether the measurements obtained fluctuated in any systematic way (eg, due to learning effect over 4 days of testing).

Test-retest reliability. Intraclass correlation coefficients were used to determine test-retest reliability. Because there were no systematic variations by day of testing for the variables investigated (see "Results" section), we obtained the necessary variance estimates for reliability assessment by assuming that the four sessions were replicates. A repeated-measures design with the four replicate measures was used (see Tab. 2 for the ANOVA table used in these estimates). Thus, using the assumption of classical test theory,[31] we assumed that only two sources of variation were possible: subject (intersubject) and variation within the subject (intrasubject).

Results

There was a wide range in the measurements for the impairment variables for this sample (Tab. 3). For example, measurements ranged from 25 to 77 degrees for cervical extension, from 122 to 174 degrees for shoulder flexion, and from 4 to 38 degrees for lumbar lordosis. Similarly, measurements for the physical performance variables varied greatly across the sample. Measurements ranged from 14.7 to 43.7 cm for functional reach, from 209.7 to 634.0 m for the 6-minute walk, and from 2.9 to 14.5 seconds for the time to move from a supine position to a standing position.

We first determined whether there were differences in the variables obtained across the different days or weeks of testing (Tab. 3). The repeated-measures ANOVA (days nested within subjects) revealed probability values of greater than .05 for all of the variables except for variables of the 360-degree turn in a standing position (time and number of steps). That is, for the most part, there were no effects of day or week of testing. With multiple ANOVAs, there is an increased probability of finding a significant result.[31] With a Bonferroni correction for multiple testing, the observed differences were not significant. Thus, the repeated-measures ANOVA revealed no effect of time, and the four data points could be used in the reliability estimates without control for days as an additional source of variation for each variable.

The estimate of reliability and the lower confidence interval for each of the measures are reported in Table 4. Test-retest reliability (ICCs) for impairment variables ranged from a low of .69 (for hamstring muscle length) to a high of .97 for lumbar flexion (Tab. 4). Test-retest reliability was .85 or better for 7 of the 13 impairment variables and .90 or better for 4 of the variables. In general, variables related to the lumbar region had the highest reliability, followed by variables of spinal configuration and muscle force. Lower-extremity ROM had the lowest reliability.

Test-retest reliability (ICCs) for physical performance variables ranged from .77 for the steps in the 360-degree turn and for the supine-to-stand task to .95 for the 6-minute walk (Tab. 4). The ICCs were greater than .85 for 3 of the 8 variables and greater than .90 for 1 variable. The lower confidence intervals were above 0.07 for 12 of the variables measured.

Table 3.
Measurements Obtained on Four Different Days, Including Means for Each Session, Univariate Statistics for the Combined Means, and Results of an Analysis of Variance

Variables	No. of Subjects	Day 1	Day 2	Day 8	Day 9	Sample[a] $\bar{X}$	SD	Range	P
Impairment variables									
Axial configuration and motion (°)									
Thoracic kyphosis	15	44.2	44.2	44.5	43.9	44.4	11.0	22–67	.38
Lumbar lordosis	15	23.5	20.9	21.4	22.2	22.4	8.0	4–38	.07
Cervical									
Flexion	15	53.3	54.1	54.7	54.9	53.8	8.8	39–75	.91
Extension	15	49.1	49.0	50.5	49.3	48.9	12.2	25–77	.67
Rotation	15	51.2	50.1	51.9	49.8	50.8	10.5	13–73	.27
Lumbar									
Flexion	15	110.5	110.7	110.6	110.1	110.1	7.6	89–124	.69
Extension	15	78.9	79.0	76.7	76.7	77.7	9.8	58–100	.30
Extremity range of motion (°)									
Shoulder flexion	15	151.0	150.7	150.6	152.3	150.4	12.2	122–174	.85
Hip flexion contracture	15	5.6	5.3	5.2	5.1	5.3	4.1	2–14	.99
Ankle dorsiflexion	15	5.8	5.6	5.4	5.2	5.4	4.8	6–14	.98
Hamstring muscle length	13	42.7	42.8	43.1	43.4	42.9	6.9	27–55	.97
Force measures (pounds of force)									
Hip abduction	15	39.1	38.0	40.9	39.2	38.3	15.7	5.7–72	.78
Ankle dorsiflexion	15	15.5	15.3	15.8	15.8	14.9	5.3	1.2–25.8	.95
Physical performance variables									
Functional axial rotation (°)	15	92.8	94.5	90.5	90.7	91.0	16.0	60–125	.51
Functional reach (cm)	14	32.3	33.3	34.3	33.0	32.5	6.6	14.7–43.7	.38
360-degree turn (s)	14	5.5	5.6	6.2	5.7	6.0	2.5	3.8–15.9	.02*
360-degree turn (steps)	14	9.0	9.2	10.8	9.4	9.5	2.9	5.5–18.5	.01*
Supine to stand (s)	14	5.4	5.9	5.7	5.3	5.6	2.2	2.9–14.5	.65
Stand to supine (s)	14	5.2	5.3	5.6	5.2	5.3	1.6	3.2–9.0	.42
6-minute walk (m)	12	477.6	455.1	468.2	478.8	461.5	94.8	209.7–634	.19
10-m walk (s)	14	8.5	8.4	8.5	8.1	8.8	2.7	5.8–17.2	.31

[a] Sample mean and standard deviation refer to the mean and standard deviation for all subjects and all test sessions; sample range refers to the range for all subjects and all test sessions.

Discussion

There is a need for measures that clinicians and researchers can choose when characterizing capabilities and limitations of patients who have PD. Such measures are required in order to monitor decline (due to the degenerative nature of the disease), to assess improvement with intervention, and to investigate the benefits of interventions. This investigation provides the first report of reliability of measurements of impairment and physical performance obtained by clinical examination for persons with mild to moderate PD. Although the sample size was small, the results provide information that can guide the choices that clinicians and researchers make regarding measurement.

Prior to establishing test-retest reliability, it was necessary to determine the effects of day or week of testing on the variables. There were few systematic variations by day or week of testing for the impairment and physical performance measurements obtained on four different days spanning a week. Subjects were always tested at the same time of day and at the same time relative to taking antiparkinsonian medications, which may have contributed to the lack of effects for day or week of testing.

Next, we determined the test-retest reliability of the measurements obtained for the variables. Reliability (ICCs) was at least .69 and is comparable to the reliability reported for these measures for subjects without

Table 4.
Intraclass Correlation Coefficients (ICCs) for Reliability of Impairment and Physical Performance Measures

Measure	No. of Subjects	ICC	Lower Confidence Interval
Axial configuration			
Kyphosis	15	.93	0.88
Lordosis	15	.87	0.78
Axial ranges			
Lumbar			
Flexion	15	.97	0.94
Extension	15	.94	0.88
Cervical			
Flexion	15	.78	0.64
Extension	15	.87	0.78
Rotation	15	.84	0.72
Extremity ranges			
Upper extremity			
Shoulder flexion	15	.80	0.67
Lower extremity			
Hip flexion	15	.72	0.56
Hamstring muscle length	13	.69	0.49
Ankle dorsiflexion	15	.80	0.66
Extremity force			
Ankle dorsiflexion	15	.91	0.84
Hip abduction	15	.88	0.78
Physical performance			
Functional axial rotation	15	.89	0.81
Functional reach	14	.84	0.71
Supine to stand	14	.77	0.61
Stand to supine	14	.80	0.45
360-degree turn			
Steps	14	.77	0.61
Time	14	.80	0.66
6-minute walk	12	.95	0.90
10-m walk	14	.87	0.77

PD.[17–21,26,27] Because the sample size was small, we also calculated the lower confidence intervals, which likewise suggested that the reliability achieved was acceptable.

There are several sources of variation contributing to the estimate of test-retest reliability: rater, within and between subjects, and instrument. Test-retest reliability for axial ROMs tended to be higher than the test-retest reliability for extremity ROM and physical performance. There is considerable room for subject variability in physical performance of tasks because of the many ways by which the tasks can be carried out. In addition, the end point of many physical performance tasks (eg, supine to standing) may be more difficult to pinpoint than the end points of measures of ROM. Variability in subject performance and rater performance might contribute to the generally lower test-retest reliability of the physical performance measures compared with the measures of axial motion. Whichever factors contribute, these measures generally have acceptable reliability, using Domholdt's criteria.[32]

Reliability estimates can be enhanced when scores from multiple trials are averaged to produce a single score (indicator) per subject.[31] The improvement in reliability has been shown to be related to the number of trials from which data are obtained using the Spearman-Brown prophesy formula.[31]

When working with subjects who have symptoms that are known to fluctuate, or when using measures for which it is difficult to define precise end points, reliability may be improved by measuring the subject on two different days and averaging the data. Using the Spearman-Brown prophesy formula,[31] we would expect a test-retest reliability (ICCs) of .87 or above for all of the physical performance variables measured in this study if data are taken on 2 days and averaged.

Although generalizability is somewhat limited by the small sample size of this study, the results indicate that acceptable reliability can be achieved for measures of limitations and physical performance for people in the early and middle stages of PD when therapists are trained to take these measurements. The measures we investigated can be used with confidence by clinicians or researchers when working with these patients. In this context, it is important to recognize that the raters underwent a training session prior to participation in the study. The complete battery of measures may be more than is required in some clinical situations. The clinician should choose those measures that are appropriate for his or her patient, given the patient's underlying impairments and the goals of intervention. Investigations currently are in progress to determine which impairment-level variables correlate with aspects of physical performance and function. In addition, we are investigating the sensitivity to change of this battery of measures.

We studied people who were in the early and middle stages of PD, who had relatively stable symptoms, and who were on a stable drug regimen in order to maximize the chance of establishing high reliability in this initial investigation. Even with rather rigorous inclusion and exclusion criteria, there was a wide range of impairment and performance measures that contributed to the success in achieving high reliability. Measures investigated in this study were chosen because of their importance for persons in the early and middle stages of PD. Investigations are needed to determine the reliability of appropriate measures of impairment and physical performance for patients who are not in a stable period with respect to symptoms of their disease and for patients who are in the later stages of PD. The reliability of these measurements obtained with larger numbers of therapists in clinical practice also needs to be determined.