2011

Functional gait assessment and balance evaluation system test: Reliability, validity, sensitivity, and specificity for identifying individuals with Parkinson disease who fall

Abigail L. Leddy  
*Washington University School of Medicine in St. Louis*

Beth Crowner  
*Washington University School of Medicine in St. Louis*

Gammon M. Earhart  
*Washington University School of Medicine in St. Louis*

Follow this and additional works at: [https://digitalcommons.wustl.edu/pt_facpubs](https://digitalcommons.wustl.edu/pt_facpubs)

**Recommended Citation**


This Article is brought to you for free and open access by the Program in Physical Therapy at Digital Commons@Becker. It has been accepted for inclusion in Physical Therapy Faculty Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Functional Gait Assessment and Balance Evaluation System Test: Reliability, Validity, Sensitivity, and Specificity for Identifying Individuals With Parkinson Disease Who Fall
Abigail L. Leddy, Beth E. Crowner and Gammon M. Earhart

Originally published online November 11, 2010

The online version of this article, along with updated information and services, can be found online at: http://ptjournal.apta.org/content/91/1/102

Collections
This article, along with others on similar topics, appears in the following collection(s):
- Balance
- Gait Disorders
- Parkinson Disease and Parkinsonian Disorders
- Tests and Measurements

e-Letters
To submit an e-Letter on this article, click here or click on "Submit a response" in the right-hand menu under "Responses" in the online version of this article.

E-mail alerts
Sign up here to receive free e-mail alerts
Functional Gait Assessment and Balance Evaluation System Test: Reliability, Validity, Sensitivity, and Specificity for Identifying Individuals With Parkinson Disease Who Fall

Abigail L. Leddy, Beth E. Crowner, Gammon M. Earhart

**Background.** Gait impairments, balance impairments, and falls are prevalent in individuals with Parkinson disease (PD). Although the Berg Balance Scale (BBS) can be considered the reference standard for the determination of fall risk, it has a noted ceiling effect. Development of ceiling-free measures that can assess balance and are good at discriminating “fallers” from “nonfallers” is needed.

**Objective.** The purpose of this study was to compare the Functional Gait Assessment (FGA) and the Balance Evaluation Systems Test (BESTest) with the BBS among individuals with PD and evaluate the tests’ reliability, validity, and discriminatory sensitivity and specificity for fallers versus nonfallers.

**Design.** This was an observational study of community-dwelling individuals with idiopathic PD.

**Methods.** The BBS, FGA, and BESTest were administered to 80 individuals with PD. Interrater reliability (n=15) was assessed by 3 raters. Test-retest reliability was based on 2 tests of participants (n=24), 2 weeks apart. Intraclass correlation coefficients (2,1) were used to calculate reliability, and Spearman correlation coefficients were used to assess validity. Cutoff points, sensitivity, and specificity were based on receiver operating characteristic plots.

**Results.** Test-retest reliability was .80 for the BBS, .91 for the FGA, and .88 for the BESTest. Interrater reliability was greater than .93 for all 3 tests. The FGA and BESTest were correlated with the BBS (r=.78 and r=.87, respectively). Cutoff scores to identify fallers were 47/56 for the BBS, 15/30 for the FGA, and 69% for the BESTest. The overall accuracy (area under the curve) for the BBS, FGA, and BESTest was .79, .80, and .85, respectively.

**Limitations.** Fall reports were retrospective.

**Conclusion.** Both the FGA and the BESTest have reliability and validity for assessing balance in individuals with PD. The BESTest is most sensitive for identifying fallers.
Parkinson disease (PD) is a neurodegenerative disease that presents a constellation of systemic motor and non–motor signs and symptoms. Postural instability, one of the most disabling cardinal signs of PD, is one of the primary reasons why someone with PD may be referred for physical therapy.\textsuperscript{1,2} Axial symptoms, such as balance impairments, are one of the main predictors of quality of life for individuals with PD and have been shown to increase fall risk.\textsuperscript{1,3} Up to 68% of individuals with PD will fall in a 1-year period, which can lead to injury and large personal and societal costs.\textsuperscript{4,5}

Ideally, individuals with PD who have impaired balance and an increased risk for falls would be identified prior to a fall and an appropriate, proactive intervention instated. At the present time, however, the best predictor of falling is a history of prior falls.\textsuperscript{4,6–9} Although many different balance assessments currently are being used, the Berg Balance Scale (BBS) can be considered a reference standard for assessing balance in people with PD, as it is one of the most commonly used balance assessments in the clinic and in research. The BBS, however, has been shown to have a ceiling effect in individuals with PD, as well as other populations.\textsuperscript{10–13} A ceiling effect occurs when the highest score on the scale does not capture or discriminate between differences in the upper end of the attribute being measured. In this case, individuals can receive a perfect score on the BBS yet still have balance impairments that need to be addressed. A balance assessment is needed for individuals with PD that can be used to: (1) accurately assess balance throughout the full ambulatory spectrum of PD and (2) identify who is at an increased risk of falling. This study investigated the properties of 2 tests, the Functional Gait Assessment (FGA) and the Balance Evaluation Systems Test (BESTest), with respect to these 2 criteria.

The FGA is an ambulation-based balance test originally proposed to assess higher-level balance in individuals with vestibular impairments.\textsuperscript{14} The precursor of the FGA was the Dynamic Gait Index (DGI), which was validated in various populations,\textsuperscript{15,16} yet had a potential ceiling effect.\textsuperscript{15–17} The FGA has excellent interrater reliability (intraclass correlation coefficient [ICC] = .93) in independently living individuals between the ages of 40 and 89 years.\textsuperscript{18} It also has acceptable interrater reliability and test-retest reliability (ICC = .86 and ICC = .74, respectively) in individuals with vestibular disorders.\textsuperscript{14} The FGA includes tasks that require many postural adjustments, as opposed to the more static items in the BBS. The reliability of this assessment for people with PD is unknown.

The assessment of balance and risk of falling in people with PD by using a combination of different measures is well supported in the literature.\textsuperscript{6,19–22} The Balance Evaluation Systems Test (BESTest) is a newly developed, multifaceted approach to assessing balance that combines portions of different balance assessments.\textsuperscript{23} Seventeen of the 36 items in the BESTest have been adopted from the following previously validated balance assessments: BBS, DGI, single-limb stance test, Timed “Up & Go” Test,\textsuperscript{24} Functional Reach Test,\textsuperscript{25} and modified Clinical Test of Sensory Interaction on Balance.\textsuperscript{26} The remaining 19 novel items include a dual-task item, postural response and compensatory stepping items, general alignment in standing, functional strength in the hips and ankles, leaning and returning to vertical, sitting on the ground and returning to a standing position, and standing on an incline. All items are divided into 6 categories (ie, “Biomechanical Constraints,” “Stability Limits and Verticality,” “Anticipatory Postural Adjustments,” “Postural Responses,” “Sensory Orientation,” and “Stability in Gait”), each representing a theoretical control system for balance.\textsuperscript{25} The categories are theorized to help guide and focus balance interventions once balance impairments are identified. The BESTest has been shown to have excellent interrater reliability (ICC = .91) and moderate validity with an individual’s self-perceived balance ($r^2 = .656$) when used to assess a mixed population of individuals with and without disease (including 3 individuals with PD).\textsuperscript{25}

To assess the validity of these new assessments, previously validated measures of balance and postural instability were used. Although the BBS is the reference standard, we compared the FGA and BESTest with this tool, as well as other correlates of postural stability and fall risk, in order to assess their possible superiority. Both disease severity and fear of falling are highly associated with postural instability and falls.\textsuperscript{7,27,28} Therefore, the modified Hoehn and Yahr scale, the Movement Disorders Society revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS), the MDS-UPDRS part 3 (motor examination) (MDS-UPDRS-3), and the Activities-specific Balance Confidence Scale (ABC) were used to assess the validity of FGA and BESTest scores.\textsuperscript{29,30}

The MDS-UPDRS total score and the Hoehn and Yahr scale are the most commonly used assessments of disease severity for individuals with PD.
The MDS-UPDRS-3 is the motor section subscale of the MDS-UPDRS that has been shown to measure motor impairment, disease severity, and disability.29,31 The ABC is a questionnaire that measures individuals’ self-perceived confidence in their balance. It has been shown to correlate with postural instability and is predictive of falls in individuals with PD.6,20

The goal of this study was to evaluate reliability, validity, sensitivity, and specificity for identifying “fallers” versus “nonfallers” for the FGA and BESTest in individuals with idiopathic PD. We hypothesized that the FGA and BESTest would be reliable assessments of balance in PD and would correlate well with the BBS, Hoehn and Yahr scale, MDS-UPDRS, MDS-UPDRS-3, and ABC. We also hypothesized that the FGA and BESTest would be able to differentiate between individuals who fall and individuals who do not fall. Furthermore, we hypothesized that the BESTest would be more sensitive and specific in determining fallers versus nonfallers because it is a more comprehensive test than either the FGA or the BBS.

Method
Participants
Eighty-two individuals with idiopathic PD were evaluated using the BBS, FGA, and BESTest. All participants met the following inclusion criteria: (1) >40 years of age, (2) diagnosed with idiopathic PD, (3) Hoehn and Yahr scale stage I to IV, (4) community dwelling, and (5) able to provide informed consent and follow instructions. Individuals were excluded from the study due to atypical parkinsonism or prior surgical management of PD (pallidotomy or deep brain stimulation).

A list of individuals with PD in the St Louis area who had been evaluated at the Washington University School of Medicine’s Movement Disorders Center, stratified by Hoehn and Yahr scale stage, was used for recruitment. Individuals were contacted using a random number generator after verifying inclusion and exclusion criteria via the database. Once contacted, the inclusion and exclusion criteria were re-verified. Individuals in the St Louis community who heard of the study from other participants or the Volunteers for Health database were allowed to participate as well. Of the 82 individuals who agreed to participate in the study, 2 were eliminated based on exclusion criteria and an unrelated illness, respectively. Twenty-five participants (31.3%) were considered to be fallers, which was defined as someone who reported 2 or more falls in the prior 6 months.

Testing Procedure
Evaluations were performed in a laboratory setting at the University of Washington School of Medicine between July and December 2009. The study was approved by the Human Research Protection Office. Participants were instructed to take their medication according to their normal regimen and were tested while on medication. After signing the approved consent forms, participants completed demographic information, reported number of falls in the prior 6 months, and took the ABC with the assistance of a caregiver as needed. The balance tests were performed in the following order: BBS, FGA, and BESTest. Each of the tests can differentiate between people with PD who have fallen and people with PD who have not fallen better than the Berg Balance Scale.

The Bottom Line
What do we already know about this topic?
People with Parkinson disease (PD) experience more falls and demonstrate increased balance deficits compared with people of the same age who do not have PD. The balance tests currently used in practice might not adequately identify these deficits and fall risk.

What new information does this study offer?
This study verifies that, because of a ceiling effect, the Berg Balance Scale, a popular balance measure, might not identify all people with PD who are at risk for falls or who have balance deficits. The Balance Evaluation Systems Test and Functional Gait Assessment—alternative tests with established reliability and validity—are now available for use in measuring balance in people with PD. Each of the tests can differentiate between people with PD who have fallen and people with PD who have not fallen better than the Berg Balance Scale.

If you’re a patient, what might these findings mean for you?
Using these new outcome measures, physical therapists may be better able to determine your fall risk.
performed the tests with shoes off, as required for the BESTest, unless they expressed discomfort without shoes. Participants were allowed to rest as needed during all portions of the evaluation. The full test session, including physical evaluations and questionnaires, required a total of approximately 2 hours to complete, with balance tests administered during the first hour. No participants indicated any wearing off of medication during the balance testing.

The BBS is a 14-item test, with each item rated from 0 (signifying poor balance) to 4 (signifying better balance). A perfect score is 56. The BBS was administered as described in the original article, with one modification. The position for the forward reach item of the BBS was altered slightly by requiring the participant to raise both arms to 90 degrees and keep the heels on the floor during the reach to further standardize the test.

The FGA is a 10-item walking-based balance test, with each item scored 0 to 3. The FGA was administered as described in the original article. It includes walking forward, backward, with eyes closed, stepping over obstacles, changing gait speeds, with different head turns, and with a narrow base of support. A higher total score signifies better balance, with a maximum score of 30.

The BESTest consists of 36 items graded on a 0 to 3 scale, with higher numbers signifying better balance and a maximum of 108 points. The total score was converted into a percentage score. The BESTest was performed as described by Horak et al., with a few slight modifications. Two trials of the compensatory stepping items were performed to encourage the participant to lean adequately and to allow the tester to adjust their pressure as needed. Only the second trial was scored. One trial, as opposed to the 2 trials in the original description, was performed for each of the items in the “Sensory Orientation” category and the functional reaches due to time constraints. Although the BESTest allows for either random number generation or counting backward by 3s in the timed get-up-and-go dual-task item, random number generation was used for all evaluations.

The 16-item ABC was administered as a questionnaire. The ABC quantifies an individual’s perceived ability to maintain his or her balance under different circumstances, using a scale of 0% (no confidence) to 100% (total confidence)

The MDS-UPDRS was followed according to Goetz et al and was administered by a trained rater. The total MDS-UPDRS score is the most common method of evaluating the severity of PD across behaviors, activities of daily living, motor abilities, and other complications of PD. The MDS-UPDRS-3 is a measure of severity of PD, as well as physical disability, and includes measures of rigidity, gait, tremor, hand/arm and leg movements (bradykinesia), speech, and facial expressions. The modified Hoehn and Yahr scale also was used to evaluate disease severity.

Reliability
Interrater reliability was determined using 3 raters (1 physical therapist student and 2 physical therapists) and a subset of 15 participants (mean MDS-UPDRS score=74.2, SD=18.6; mean disease duration=6.8 years, SD=3.38; 21% [n=5] fallers; Hoehn and Yahr scale stage 1=2, stage 2=1, stage 3=6, stage 3=3, and stage 4=2) twice with a 2-week (range=11-16 days) intervening period. For both evaluations, participants were tested at the same time of day and instructed to take medications as usual to reduce the likelihood of on/off medication fluctuations causing variations in PD signs and symptoms. Test-retest reliability was determined for a physical therapist and the physical therapist student.

Once the reliability for all tests was determined to be good using the initial subset of participants, the remaining balance evaluations for assessing validity, sensitivity, and specificity for detecting fallers versus
nonfallers were performed by the physical therapist student.

**Data Analysis**
Statistics were calculated using SPSS for Windows (version 17.0).* Independent-sample *t* tests and Mann-Whitney *U* tests were used to compare age, disease duration, and disease severity of the fallers and nonfallers. Interrater and test-retest reliability were calculated using ICC (2,1) because raters tested participants using information that any physical therapist or physical therapist student would have available. To quantify validity, the correlations between the 2 newer tests and commonly used measures of PD severity and balance were used. The FGA and BESTest were compared with the BBS, ABC, MDS-UPDRS, MDS-UPDRS-3, and Hoehn and Yahr scale stages using Spearman correlation coefficients. For all Spearman correlations, .00 to .25 = little to no relationship, .25 to .50 = fair, .50 to .75 = moderate, and .75 to 1.00 = high correlation. To maintain an alpha level of .05, a Bonferroni correction for multiple comparisons required a *P* value of .002. Receiver operating characteristic (ROC) plots were made for each test, and cutoff values were chosen by selecting the score with the minimal value of:

\[
(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2,
\]

which maximizes sensitivity and specificity. To allow a more accurate comparison with prior studies identifying individuals with PD who were fallers, an alternative cutoff point also was calculated by maximizing sensitivity and minimizing negative likelihood (LR−) ratio. Post-test probabilities were calculated using LR−, positive likelihood ratios (LR+), and the pretest probability of falling from this study sample to allow the balance test results to be interpreted more completely. Overall accuracy for each balance test was assessed using the area under the curve (AUC). The AUC is the probability that the individual with PD who is a faller will be correctly identified, given 2 randomly selected individuals are chosen, 1 who is a faller and 1 who is not a faller. For AUC analysis, 0.5 = test due to chance, 0.5 to 0.7 = low accuracy, 0.7 to 0.9 = moderate accuracy, 0.9 to 1.0 = high accuracy, and 1.0 = a perfect test. Sample size calculations for the study were based on a power of 0.80 and an alpha level of .05. Interrater reliability based on 3 raters, a null ICC of .50, and an acceptable reliability of .80 required 15 participants. Test-retest reliability based on 2 trials of testing required 22 participants. Eighty-one participants were required for sensitivity, specificity, and ROC curves by estimating a 30% faller rate, a confidence interval width of 0.20, and a 95% confidence level.

**Role of the Funding Source**
This work was directly funded by a grant from the Davis Phinney Foundation.

---

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

---

**Table 1.**
Participant Demographic and Disease Severity Information Overall and Separated Out by Fallers and Nonfallers

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall (N=80)</th>
<th>Fallers (n=25)</th>
<th>Nonfallers (n=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (%)</td>
<td>59</td>
<td>64</td>
<td>56</td>
</tr>
<tr>
<td>Age (y)</td>
<td>68.2</td>
<td>68.8</td>
<td>67.9</td>
</tr>
<tr>
<td>SD</td>
<td>9.3</td>
<td>7.8</td>
<td>10.0</td>
</tr>
<tr>
<td>Range</td>
<td>45–88</td>
<td>55–88</td>
<td>45–85</td>
</tr>
<tr>
<td>Disease duration (y)</td>
<td>8.5</td>
<td>11.4</td>
<td>7.15</td>
</tr>
<tr>
<td>SD</td>
<td>0.54</td>
<td>5.5*</td>
<td>3.81</td>
</tr>
<tr>
<td>Range</td>
<td>1–25</td>
<td>4–25</td>
<td>1–16</td>
</tr>
<tr>
<td>MDS-UPDRS* total score</td>
<td>72.6</td>
<td>93.8</td>
<td>62.9</td>
</tr>
<tr>
<td>SD</td>
<td>25.1</td>
<td>23.1*</td>
<td>19.0</td>
</tr>
<tr>
<td>Range</td>
<td>25–135</td>
<td>46–135</td>
<td>25–103</td>
</tr>
<tr>
<td>Hoehn and Yahr scale stage</td>
<td>2.45</td>
<td>2.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Median</td>
<td>2.5</td>
<td>3.0*</td>
<td>2.0</td>
</tr>
<tr>
<td>Stage 1</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Stage 2</td>
<td>27</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Stage 2.5</td>
<td>30</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Stage 3</td>
<td>13</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Stage 4</td>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

* Statistically significant difference between fallers and nonfallers (*P* <.001).
* MDS-UPDRS = Movement Disorders Society revision of the Unified Parkinson’s Disease Rating Scale.
ulation and grant UL1 RR024992 and sub-award TL1 RR024995 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. The funding source had no impact or input on the design, conduct, or reporting of this study.

**Results**

The subgroup of fallers was statistically different from the nonfallers in disease duration, MDS-UPDRS scores, and Hoehn and Yahr scale staging, with the fallers having more advanced PD (Tab. 1). The BBS scores were significantly left skewed (Kolmogorov-Smirnov test, \( P = .035 \)), with 10% (\( n = 8 \)) having a perfect score (including 1 faller) and 45% (\( n = 36 \)) scoring within the top 10% of the test (including 5 fallers). The same 8 individuals who received a perfect score on the BBS showed varying levels of balance impairments on the FGA and BESTest, having scores ranging from 23 to 30 on the FGA and from 82% to 94% on the BESTest. The test scores were normally distributed for the FGA and the BESTest. Only 1.3% (\( n = 1 \)) scored perfectly on the FGA, with 13% (\( n = 10 \)) within the top 10% of the test, none of whom were fallers. There were no perfect scores on the BESTest, with only 6.4% (\( n = 5 \)) scoring in the top 10%, none of whom were fallers (Fig. 1).

Interrater reliability among the 3 raters was excellent and comparable for all 3 tests, with ICCs greater than .93 (Tab. 2). Test-retest reliability was similar between the physical therapist student and the licensed physical therapist for both the BBS, which showed moderate reliability, and the BESTest, which had high reliability. There was a larger discrepancy between the test-retest scores for the FGA, with the physical therapist having excellent reliability, but the physical therapist student having only moderately good reliability. The BESTest had the highest test-retest reliability of the 3 assessments (Tab. 2).

The FGA was highly correlated with the BBS and moderately correlated with the ABC and disease severity measures, including MDS-UPDRS, MDS-UPDRS-3, and Hoehn and Yahr scale stage. The BESTest was highly correlated with the BBS, ABC, MDS-UPDRS, and MDS-UPDRS-3. It was

---

**Figure 1.**

Distribution of scores for the Berg Balance Scale (BBS), Functional Gait Assessment (FGA), and Balance Evaluation Systems Test (BESTest). K-S = one-sample Kolmogorov-Smirnov Test.
FGA and BESTest in Individuals With PD

Table 2.
Interrater and Test-Retest Reliability for the Berg Balance Scale (BBS), Functional Gait Assessment (FGA), and Balance Evaluation Systems Test (BESTest)*

<table>
<thead>
<tr>
<th>Type of Reliability</th>
<th>Balance Assessment Measure</th>
<th>Testers</th>
<th>ICC (2,1)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interrater reliability (n=15)</td>
<td>BBS</td>
<td>2 PTs, 1 SPT</td>
<td>.95</td>
<td>.88–.98</td>
</tr>
<tr>
<td></td>
<td>FGA</td>
<td>2 PTs, 1 SPT</td>
<td>.93</td>
<td>.84–.98</td>
</tr>
<tr>
<td></td>
<td>BESTest</td>
<td>2 PTs, 1 SPT</td>
<td>.96</td>
<td>.89–.99</td>
</tr>
<tr>
<td>Test-retest reliability (n=24)</td>
<td>BBS</td>
<td>SPT</td>
<td>.79</td>
<td>.57–.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT</td>
<td>.80</td>
<td>.60–.91</td>
</tr>
<tr>
<td></td>
<td>FGA</td>
<td>SPT</td>
<td>.80</td>
<td>.58–.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT</td>
<td>.91</td>
<td>.80–.96</td>
</tr>
<tr>
<td></td>
<td>BESTest</td>
<td>SPT</td>
<td>.91</td>
<td>.80–.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT (n=23)</td>
<td>.88</td>
<td>.72–.95</td>
</tr>
</tbody>
</table>

*All intraclass correlation coefficients (ICCs) were significant, with \( P \) values of <.01. PT = physical therapist, SPT = physical therapist student.

Table 3.
Validity (Spearman \( \rho \)) of the Functional Gait Assessment (FGA) and Balance Evaluation Systems Test (BESTest) With Measures of Disease Severity (Hoehn and Yahr Scale, Movement Disorders Society Revision of the Unified Parkinson’s Disease Rating Scale [MDS-UPDRS], and Movement Disorders Society Revision of the Unified Parkinson’s Disease Rating Scale Part 3 [Motor Examination] [MDS-UPDRS-3]), Self-Perceived Balance (Activities-specific Confidence Scale [ABC]), and the Balance Reference Standard (Berg Balance Scale [BBS])*

<table>
<thead>
<tr>
<th></th>
<th>Disease Severity</th>
<th>Self-Perceived Balance</th>
<th>Balance Reference Standard</th>
<th>New Balance Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modified Hoehn and Yahr Scale</td>
<td>MDS-UPDRS-3 Score</td>
<td>MDS-UPDRS Score</td>
<td>ABC Score</td>
</tr>
<tr>
<td>Self-Perceived Balance</td>
<td>ABC Score (N=80)</td>
<td>(-.591)</td>
<td>(-.523)</td>
<td>(-.726)</td>
</tr>
<tr>
<td>Balance Reference Standard</td>
<td>BBS Score (N=80)</td>
<td>(-.629)</td>
<td>(-.710)</td>
<td>(-.710)</td>
</tr>
<tr>
<td>New Balance Measures</td>
<td>FGA Score (n=79)</td>
<td>(-.670)</td>
<td>(-.669)</td>
<td>(-.692)</td>
</tr>
<tr>
<td>BESTest Score (n=79)</td>
<td>(-.736)</td>
<td>(-.758)</td>
<td>(-.780)</td>
<td>.757</td>
</tr>
</tbody>
</table>

*All correlations were significant at \( P < .001 \) (2-tailed).

moderately correlated with Hoehn and Yahr scale stage (Tab. 3).

Receiver operating characteristic plots for the BBS, FGA, and BESTest are shown in Figure 2. The cutoff scores suggested based on maximizing both sensitivity and specificity were 47/56 for the BBS, 15/30 for the FGA, and 69% for the BESTest (Tab. 4). Based on these cutoff points, similar specificity was found for all 3 tests, with sensitivity being highest for the BESTest. The BESTest had the highest LR+ and the lowest LR−. The pretest probability of being a faller in this study sample was 31%. The posttest probability for an individual to be a true faller based on these cutoff scores ranged from 36% for the FGA to 44% for the BBS, given a positive test (identified as a faller), and from 7% on the BBS to 0% on the FGA or BESTest, given a negative test (identified as a non-faller) (Tab. 4).

When sensitivity was maximized and LR− was minimized, the specificity decreased for all of the tests (0.19–0.47), and sensitivity increased to 9% on the BESTest, given a negative test (identified as a non-faller) (Tab. 4).

The nonparametric AUC was calculated because the BBS scores were not normally distributed. The AUC was highest for the BESTest, though the 95% confidence interval overlapped among the 3 tests (Tab. 4). All 3 tests showed moderate accuracy. There were no adverse events during the course of the study.

Discussion
Measures that can be used to assess balance across the ambulatory spectrum of PD and decipher who is and is not at risk for falling are of vital
importance to determine appropriate treatment for individuals with PD. Early interventions may prevent or decrease the negative effects of postural instability. This study demonstrates that both the FGA and BESTest are reliable and valid measures of balance in PD, with acceptable sensitivity and specificity for identifying fallers versus nonfallers.

Reliability
The high interrater reliability for the BBS, FGA, and BESTest found in this study should be generalizable, as training for each test included only resources available in the clinical setting. The video training for the BESTest was used and is available for purchase online. The interrater reliability was comparable to what has been reported for the FGA and BESTest in prior studies with other populations.

Test-retest reliability quantifies the difference between 2 test sessions, including variability in the tester’s rating and the participant’s performance. Test-retest reliability was similar for the physical therapist (ICC = .80) and the physical therapist student (ICC = .79) for the BBS, yet lower than in prior studies. Steffen and Seney and Lim et al reported ICCs of .94 and .87, respectively, in individuals with PD. One reason for this difference might be the longer time between test sessions (2 weeks) in our study. Variations in performance due to medication schedule.

Table 4.
Cutoff Points With Associated Sensitivity, Specificity, Likelihood Ratios, Posttest Probabilities, and Area Under the Curve (AUC) of the Receiver Operating Characteristic Plot for the Berg Balance Scale (BBS), Functional Gait Assessment (FGA), and Balance Evaluation Systems Test (BESTest)*

<table>
<thead>
<tr>
<th>Balance Assessment Measure</th>
<th>AUC (95% CI)</th>
<th>Score</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>LR+ (95% CI)</th>
<th>LR– (95% CI)</th>
<th>Posttest Probability With Test ≤ Cutoff Value</th>
<th>Posttest Probability With Test &gt; Cutoff Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS</td>
<td>0.79 (0.68–0.91)</td>
<td>≤47/56*</td>
<td>0.72</td>
<td>0.75</td>
<td>2.83 (1.69–4.73)</td>
<td>0.38 (0.20–0.72)</td>
<td>56.3%</td>
<td>14.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≤52/56</td>
<td>0.92</td>
<td>0.47</td>
<td>1.74 (1.32–2.30)</td>
<td>0.17 (0.04–0.66)</td>
<td>44.2%</td>
<td>7.1%</td>
</tr>
<tr>
<td>FGA</td>
<td>0.80 (0.69–0.90)</td>
<td>≤15/30*</td>
<td>0.72</td>
<td>0.78</td>
<td>3.24 (1.86–5.65)</td>
<td>0.36 (0.19–0.69)</td>
<td>59.6%</td>
<td>14.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≤27/30</td>
<td>1.00</td>
<td>0.19</td>
<td>1.23 (1.08–1.39)</td>
<td>0.00 (unable to calculate)</td>
<td>35.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>BESTest</td>
<td>0.85 (0.77–0.94)</td>
<td>≤69%*</td>
<td>0.84</td>
<td>0.76</td>
<td>3.49 (2.11–5.77)</td>
<td>0.21 (0.09–0.52)</td>
<td>61.3%</td>
<td>8.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≤84%</td>
<td>1.00</td>
<td>0.39</td>
<td>1.64 (1.32–2.02)</td>
<td>0.00 (unable to calculate)</td>
<td>42.7%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

* For all balance tests, 2 cutoff values are reported. The first cutoff value (indicated by asterisk) was chosen to maximize both sensitivity and specificity. The second cutoff value was chosen by maximizing sensitivity and negative likelihood ratio (LR–). Pretest probability for being a faller was 31.3%. CI = confidence interval, LR+ = positive likelihood ratio.
were controlled for by keeping testing at the same time of day; however, it is possible that participants varied slightly in their medication regimen, causing changes in individual performances.

Test-retest reliability for the BESTest was high and similar for the physical therapist and the physical therapist student. Despite its complexity and greater number of items, the BESTest can be used successfully and consistently with available training tools. This is the first study to report test-retest reliability for the BESTest.

Validity
The BESTest had the best validity, with higher correlation to all measures of disease severity and self-perceived balance than the BBS and FGA. The highest correlation, though, was between the BBS and the BESTest, signifying that balance is still the primary construct being assessed. The FGA had moderate overall validity; its highest correlation was with the BBS. Both the FGA and the BESTest are valid measures of balance in individuals with PD.

Ceiling Effect in the BBS
Although the BBS is a good assessment of balance in other populations and has served as a functional balance and fall risk measure in the past, it does not challenge balance sufficiently to allow detection of balance impairments across the full disease spectrum of PD. The BBS does include standing up, turning around, and bending over, which are 3 of the most common ways falls occur in people with PD. However, unlike the situation where falls occur, the items allow full attention to be allocated to these simple tasks, possibly missing those who would lose their balance under nontested circumstances. It has been shown that attention allocation can drastically change balance deficits in individuals with PD, whether it is a dual-task situation or the individual perceives that he or she being observed. The participants were fully focused, and many were able to successfully complete the BBS items, although these same tasks are problematic under normal circumstances. Bradykinetic and improperly scaled postural adjustments also are implicated for loss of balance in people with PD, which many of the static items in the BBS do not require or require at minimal levels only. Due to these missing components in the test, the ceiling effect noted in previous studies is quite severe. Twenty-five percent (n=5) of the fallers in this study scored within the top 10% of the BBS, and 4% of fallers (n=1) received a perfect score. The BBS is unable to identify some individuals with PD who are at risk for falls, let alone identify the more subtle balance deficits that occur in the disease prior to the occurrence of falls. According to Steffen and Seney, the BBS score must change at least 5 points to show a true change in balance. Therefore, 43% of the individuals in our study (those scoring ≥52/56) would not even be able to show any balance progress using the BBS. This lack of ability to identify early balance impairments in individuals with PD may prevent important early intervention measures.

Identifying Fallers and Nonfallers
We present cutoff scores for making distinctions between fallers and nonfallers using 2 methods. The first method maximizes both sensitivity and specificity, minimizing both false-positive and false-negative identifications of fallers. The second method maximizes sensitivity and minimizes LR-. This second method has been included to allow comparison with prior cutoff scores reported in the literature. Both methods maximize sensitivity, as having high sensitivity is important for deciding who is at risk for falls. However, there is a trade-off between sensitivity and specificity, as shown by the large decrease in specificity when the second method was used. High sensitivity paired with low specificity causes an increase in the false-positive rate. While the harm of false negatives is apparent (ie, missing someone who was at risk for falling), there is perhaps equal harm in false positives, as inappropriately treating an individual who is not truly at risk for falling can be costly to society, an inappropriate use of the patient’s limited financial resources, and a drain on the individual’s (as well as the caregiver’s) time and energy.

When discriminating who is a faller, the suggested cutoff value of 47/56 for the BBS was chosen based on maximized sensitivity and specificity. This cutoff is between the points chosen in prior studies. A cutoff value of 44/56 was suggested by Landers et al by maximizing the change in positive posttest probability, which allows for a higher false-negative rate. Dibble and Lange chose 54/56 by maximizing sensitivity and minimizing LR-. Which allows for a higher false-positive rate. Applying the same method used by Dibble and Lange to the present data would yield a cutoff of 52/56. Differences in definition of fallers and in fall rates, as well as mode of recruitment, could contribute to the differences seen between the present results and those of Dibble and Lange. Dibble and Lange defined a faller as someone having 2 or more falls in a 12-month period, their study had a 51% fall rate, and all participants had been referred for physical therapy. Our study included participants from the community who were not actively seeking treatment and, theoretically, less impaired.

The FGA cutoff value for identifying fallers was 15/30 when maximizing sensitivity and specificity. No ceiling effect seen for the test, showing that...
the changes made to the DGI to create the FGA were effective. The sensitivity and specificity for identifying fallers in our study were 0.72 and 0.78, respectively.

When maximizing sensitivity and specificity, the BESTest had the highest accuracy in the 3 tests without compromising specificity. Based on the current study, the cutoff score for identifying fallers among individuals with PD is 69%. The overall predictive accuracy was highest for the BESTest based on the AUC, although only slightly higher than for the BBS and FGA. The higher AUC indicates a higher probability that those who are fallers will have a worse balance score than those who are not fallers. Eighty-five out of 100 times, the BESTest score of a randomly selected individual who is a faller will be lower than the score of a randomly selected individual who is a nonfaller.

It is important to understand that the purpose of cutoff scores is to help identify those who are at greater risk clinically for falling. However, they are not to be used as a true dichotomous scale. Those near the cutoff scores are at more risk than those who are farther away (higher scores). The cutoff values presented are to be used by clinicians to help understand the scores on these tests and the balance impairment level of those who fall normally compared with those who do not fall.

The LR can be interpreted as how many times more likely a faller is to receive a score below the cutoff than a nonfaller. For example, the BESTest had the highest LR+, with a faller being 3.5 times more likely to score ≤69% than a nonfaller. The BESTest had the lowest LR−, with fallers 2.1 times more likely to have a score above 69% than nonfallers. All posttest probabilities should be compared with the pretest probability of 31% in this study.

To assess use of these cutoff scores in a clinical population with higher disease severity, a sensitivity analysis was completed, identifying fallers in individuals with Hoehn and Yahr scale stages 2.5 to 4.0 only (faller rate = 47%). In all 3 tests, the cutoff values chosen using both previously mentioned methods were the same as or slightly lower than what is presented using all study participants Hoehn and Yahr scale stages 1 to 4, except for the BBS, which suggested a cutoff value of 55/56 for maximizing sensitivity and minimizing LR−.

Although other literature suggests that a combination of tests should be used to assess balance in individuals with PD, the recommendations are highly variable. Lim et al suggested the combination of the UPDRS, UPDRS-3, Timed “Up & Go” Test, Functional Reach Test, and timed 10-meter walk for assessing balance in the home. Dibble et al proposed the use of multiple tests to increase the posttest probability of accurately assessing who was at risk for falling. They used prior reported cutoff values of the Functional Reach test, Dynamic Gait Index, BBS, and Timed “Up & Go” Test and then determined fall risk based on the number of tests on which the individual scored below the cutoff value. The combination suggested by Mak and Pang to identify recurrent fallers and nonrecurrent fallers includes only the UPDRS and ABC scores. Each study agrees that a combination of assessments is necessary due to the multifaceted structure of balance, although all the resulting solutions are different.

The BESTest is the only test, or combination of tests, that has different items categorized according to the theoretical control systems of balance. This design might allow the identification of fallers, as well as help determine the main contributor to the underlying balance deficit. A specified combination of items, as presented in the BESTest, would allow standardization of clinical and research evaluations to enable comparisons to be made, as opposed to having each therapist or researcher pick from an assortment of balance tests.

Limitations

One limitation of the study is the use of a retrospective fall report. However, Bloem et al found a similar rate of falling (25%) in a 6-month prospective study of falls in individuals with PD. Another limitation is that the balance tests were not randomized. This limitation did not seem to affect the outcome of the study, as all participants finished the tests, and during the reliability subset, the BESTest was highly reliable, although it was administered toward the end of testing. Participants also were allowed to rest as often as they wanted to prevent any fatigue.

For reliability testing, only one rater actually performed the balance tests, while the other raters concurrently observed. Although there is a written script for administering the tests, the possible differences among testers’ verbal and nonverbal communication may have contributed to increased variability in participant performance had the raters each administered the tests separately. For all testing, the raters were blinded only to fall status.

Summary and Implications

Future studies should focus on whether the FGA and BESTest are predictive of falls using the cutoff points described here in a prospective manner. The responsiveness of these balance tests to assess change in a single individual over time (ie, clinically significant differences in scores) also is important to explore for better clinical discrimination of effectiveness of different balance interventions. A shortened version of the BESTest that is specific
for individuals with PD might increase its use in the clinical setting.

In summary, the BBS has a ceiling effect and thus may not adequately assess balance in the early stages of postural instability in individuals with PD. Both the FGA and the BESTest are reliable and valid measures of balance that can be used throughout PD Hoehn and Yahr scale stages 1 to 4. Both tests can be used to discriminate between fallers and nonfallers. The BESTest has the highest sensitivity for identifying who is a faller; however, it takes longer to administer than the FGA or BBS.

All authors provided concept/idea/research design and data collection. Ms Leddy provided writing, data analysis, and clerical support. Ms Leddy and Dr Earhart provided project management and fund procurement. Dr Crowner and Dr Earhart provided participants and consultation (including review of manuscript before submission). Dr Earhart provided facilities/equipment. Special thanks go to all of the research participants for their time, as well as to Ryan Duncan, John Michael Rotello, and Vanessa Heiland Chapdelaine for their help with data collection.

The study was approved by the Human Research Protection Office of Washington University School of Medicine.

This research was presented in abstract/poster format at the Missouri Physical Therapy Association Spring Conference; April 16–18, 2010; St Louis, Missouri; at the National Predoctoral Clinical Research Training Program Meeting; May 3–4, 2010; St Louis, Missouri; and at the World Parkinson Congress; September 28–October 1, 2010; Glasgow, Scotland.

This publication was directly funded by the Davis Phinney Foundation and grant UL1 RR024992 and sub-award TL1 RR024995 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Additional support was provided by the Greater St Louis Chapter of the American Parkinson Disease Association (APDA) and the APDA Center for Advanced PD Research at Washington University. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the Davis Phinney Foundation, NCRR, NIH, or the APDA.

This article was submitted March 31, 2010, and was accepted August 15, 2010.


References


Functional Gait Assessment and Balance Evaluation System Test: Reliability, Validity, Sensitivity, and Specificity for Identifying Individuals With Parkinson Disease Who Fall
Abigail L. Leddy, Beth E. Crowner and Gammon M. Earhart

Originally published online November 11, 2010