Test-Retest Reliability, Validity, and Minimal Detectable Change of the Balance Evaluation Systems Test to Assess Balance in Persons with Multiple Sclerosis

Kathryn D. Mitchell, PT, DPT, NCS, MSCS; Han Chen, MD, MPH; Sheri P. Silfies, PT, PhD

Background: Individuals with multiple sclerosis (MS) have balance deficits that result in falls. Balance tests developed for older adults do not discriminate between fallers and nonfallers with MS. The Balance Evaluation Systems Test (BESTest) identifies body systems contributing to imbalance/fall risk. This study evaluated the test-retest reliability and construct validity of the BESTest to assess balance in individuals with MS, compared the diagnostic accuracy of the BESTest and Berg Balance Scale (BBS), and investigated the minimal detectable change of the BESTest in this population.

Methods: Twenty individuals with MS were recruited. Age, onset of disease, and fall history were obtained. The following measures were used: Expanded Disability Status Scale, University of Illinois at Chicago Fear of Falling Measure, BESTest, and BBS. Twelve participants were retested to determine test-retest reliability of the BESTest.

Results: The BESTest demonstrated strong test-retest reliability (intraclass correlation coefficient [3,1] = 0.98 [95% CI, 0.91-0.99], P = .01) and was highly correlated with the BBS (r = 0.94, P = .01). The BBS had poor accuracy to identify fallers. Using a score of 81%, the BESTest had moderate sensitivity (0.89) to identify fallers. The standard error of the measurement for the BESTest was 4.16 points, which translates into a minimal detectable change of 9.7% [90% CI] to 11.5% [95% CI].

Conclusions: The BESTest demonstrates excellent test-retest reliability and diagnostic utility as a measure of functional balance in persons with moderate MS. A change in BESTest score of at least 10% may indicate a true change in balance performance in this population. Int J MS Care. 2018;20:231-237.
bilitation of persons with MS who have fallen or who are at risk of falling.1,4-6

Physical therapists routinely use balance tests designed for assessment of balance in older adults to assess balance in the MS population. Despite the frequency of use, the Berg Balance Scale (BBS),7 the Timed Up and Go test,8 and the Dynamic Gait Index9 do not discriminate between fallers and nonfallers with MS.10 The limitation of the BBS, the most frequently used clinical tool, is that it does not assess balance while walking, and its functional items do not challenge the dynamic postural control of a younger, more active person.10,11 The 14-item BBS assesses quiet stance with a narrow base of support and standing with eyes closed, but it does not examine the impact of absent vision and reduced somatosensory input on postural responses. In addition, the BBS does not examine reactive postural control, a significant challenge for persons with MS.4 As a result, use of the BBS as a fall risk measure can be restricted by a ceiling effect.10,11 A ceiling effect occurs when a score on the upper end of a functional measure does not accurately capture performance at the high end of the construct being tested.12 In this case, the BBS does not appropriately challenge the balance abilities of persons with MS, leading to an underassessment of actual fall risk. In addition, it does not identify specific postural control impairments unique to persons with MS.10,11

The BESTest13 is a newer functional balance test developed to identify impairments in body systems that contribute to imbalance and/or fall risk. The tool was created to address the decreased sensitivity (ability to identify the impairment) of the existing clinical assessments and to assist in the development of focused interventions targeting the root cause of balance impairments. The BESTest consists of 36 items divided into six subsections, each designed to test a different dimension of postural control and balance: biomechanical constraints, limits of stability, sensory orientation in standing, anticipatory postural responses, reactive postural responses, and gait.13 One of the advantages of the BESTest is that the items include a variety of difficulty levels, making it resistant to a ceiling effect with higher-functioning patients.12,14,15

The BESTest has been validated for use in several patient groups that also have increased fall risks, including those with Parkinson’s disease, subacute stroke, and vestibular dysfunction.12,13,16,17 Its use in the MS population has been limited. The validity of the BESTest to identify balance impairments and fall history has been studied in a small group of persons with mild MS disability.18 Secondary analysis of data from that study also validated the Mini-BESTest and the Brief-BESTest in an MS population with mild disability.19 Recent work by Potter et al.20 examined reliability and validity in 21 persons with mild MS-related disability and found that the BESTest exhibited strong test-retest reliability (0.94) and internal consistency (0.97). The reliability and diagnostic accuracy of the BESTest in persons with moderate MS-related disability (Expanded Disability Status Scale [EDSS] scores of 3.0-6.0)11 has not been investigated.

We undertook this study to evaluate the reliability and validity of the BESTest in a group of individuals with moderate MS-related functional impairments. Specifically, the purposes of this study were 1) to evaluate the test-retest reliability and construct validity of the BESTest to assess balance in individuals with MS with moderate disability, 2) to compare the diagnostic accuracy of the BESTest and the BBS to identify individuals with MS and a history of falls, and 3) to determine the minimal detectable change (MDC) of the BESTest. We hypothesized that the BESTest would demonstrate excellent test-retest reliability and moderate accuracy for diagnosing falls and present as a valid balance assessment tool with small measurement error in individuals with moderate MS.

Methods
Participants
The participants were 20 ambulatory adults (Table 1) recruited from an urban university-based MS clinic. They had a confirmed diagnosis of MS and were capable

### Table 1. Characteristics of the 20 study participants

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female sex, No. (%)</td>
<td>16 (80)</td>
</tr>
<tr>
<td>Fallers (&gt;1 fall in past month), No. (%)</td>
<td>9 (45)</td>
</tr>
<tr>
<td>Age, y</td>
<td>43.3 ± 10.2 (24-61)</td>
</tr>
<tr>
<td>Time since diagnosis, y</td>
<td>11.2 ± 7.7 (1-25)</td>
</tr>
<tr>
<td>EDSS score</td>
<td>4.8 ± 1.8 (2.0-6.5)</td>
</tr>
<tr>
<td>UIC FFM score</td>
<td>34.15 ± 8.3 (21-48)</td>
</tr>
<tr>
<td>BESTest total score, %</td>
<td>71.1 ± 20.8 (32-97)</td>
</tr>
<tr>
<td>BBS score</td>
<td>47.7 ± 11.4 (15-56)</td>
</tr>
</tbody>
</table>

Note: Unless otherwise indicated, values are given as mean ± SD (range).
Abbreviations: BBS, Berg Balance Scale; BESTest, Balance Evaluation Systems Test; EDSS, Expanded Disability Status Scale; UIC FFM, University of Illinois at Chicago Fear of Falling Measure.
of walking household distances with or without an assistive device. Exclusion criteria were inability to walk household distances and the presence of other neurologic diagnoses. Individuals were invited to participate when on site for their regular clinic appointments.

**Procedure**

Consented participants completed a demographic and fall history questionnaire and the University of Illinois at Chicago Fear of Falling Measure (UIC FFM). Disease severity was assessed using the EDSS. Functional balance of each participant was assessed using the BESTest and the BBS. Both tests were administered in one session, with the order of balance tests randomized. To account for fatigue, a 10-minute break was taken between the administrations of the two functional balance tests. Available participants were retested 7 to 14 days later to examine the test-retest reliability of the BESTest. The results of all testing were reviewed with the participant after completion of the BESTest retest. One physical therapist (K.D.M.), a board-certified neurologic clinical specialist and MS-certified specialist trained in the use of both tools, completed all testing. The study was approved by the Drexel University institutional review board in accordance with the principles expressed in the World Medical Association Declaration of Helsinki (October 2013 version).

**Outcome Measures**

Demographic information collected included age, sex, MS subtype (relapsing-remitting, secondary progressive, primary progressive), and years since MS diagnosis. The participants also self-reported the number of falls sustained in the month before the assessment. We chose to use a 1-month recall secondary to functional fluctuations associated with MS and the high frequency of falls reported in the MS population seen in our specialty clinic. Disease severity was captured by a trained clinician (K.D.M.) during the first visit using the EDSS. The EDSS is a comprehensive assessment of MS disease-related disability that quantifies disease state using an ordinal scale ranging from 0 to 10, in 0.5 increments, across eight systems, with a larger score indicating greater disability. The EDSS scores of 3.0 to 6.0 indicate moderate disease-related disability.

To measure fear of falling, the UIC FFM was administered. The UIC FFM is a self-report questionnaire designed to measure an older adult’s fear of falling. It asks the respondent to rate comfort with performing 16 common activities (eg, walking carrying items, standing on a step stool, standing on a moving bus), with scores ranging from 16 to 48 and higher scores representing less fear of falling. Using a cutoff score of 37 or less, the UIC FFM correctly identifies older adult fallers with a sensitivity of 0.64 and a specificity of 0.80. Use of this instrument with persons with MS has not been reported.

The BBS is a 56-point, 14-item functional balance assessment tool designed to identify fall risk in community-dwelling older adults. It is considered the reference standard, and its use as a functional balance tool for individuals with neurologic conditions has been well documented. The BBS is scored on a 5-point ordinal scale (0-4), with a higher score indicating better performance. Multiple studies have reported strong test-retest reliability (ICC = 0.85-0.96 [95% CI, 0.72-0.98]) and interrater reliability (ICC = 0.96-0.99 [95% CI, 0.99-1.00]) for the BBS. Despite its ease of use and popularity, the BBS has been shown to have a ceiling effect when used to identify fallers in the MS population.

The BESTest consists of 36-items divided into six subsections, each designed to test a different dimension of postural control: biomechanical constraints (eg, ankle range of motion and strength), limits of stability (eg, forward and lateral functional reach), sensory orientation in standing (eg, standing on foam with eyes open and eyes closed), anticipatory postural responses (eg, sternal nudge, alternating stair tap), reactive postural responses (eg, compensatory stepping), and gait (eg, changing gait speed, walking with head turns). Each item is scored from 0 (severe impairment) to 3 (no impairment), and the total score is expressed as a percentage of points scored of 108 total points.

**Statistical Analysis**

Statistics were calculated using a statistical software program (IBM SPSS Statistics for Windows, version 23.0; IBM Corp, Armonk, NY). Baseline and demographic characteristics are summarized using minimal standard, and its use as a functional balance tool for individuals with neurologic conditions has been well documented.13 Their use as a functional balance tool for individuals with neurologic conditions has been well documented.13,18,24 The BBS is scored on a 5-point ordinal scale (0-4), with a higher score indicating better performance. Multiple studies have reported strong test-retest reliability (ICC = 0.85-0.96 [95% CI, 0.72-0.98]) and interrater reliability (ICC = 0.96-0.99 [95% CI, 0.99-1.00]) for the BBS. Despite its ease of use and popularity, the BBS has been shown to have a ceiling effect when used to identify fallers in the MS population.

The BESTest consists of 36-items divided into six subsections, each designed to test a different dimension of postural control: biomechanical constraints (eg, ankle range of motion and strength), limits of stability (eg, forward and lateral functional reach), sensory orientation in standing (eg, standing on foam with eyes open and eyes closed), anticipatory postural responses (eg, sternal nudge, alternating stair tap), reactive postural responses (eg, compensatory stepping), and gait (eg, changing gait speed, walking with head turns). Each item is scored from 0 (severe impairment) to 3 (no impairment), and the total score is expressed as a percentage of points scored of 108 total points.

**Statistical Analysis**

Statistics were calculated using a statistical software program (IBM SPSS Statistics for Windows, version 23.0; IBM Corp, Armonk, NY). Baseline and demographic characteristics are summarized using minimal and maximal variables, means, and SDs for continuous variables and the number and percentage of individuals in each category for categorical variables. Floor and ceiling effects were analyzed as the percentage of participants who received the minimum and maximum possible scores for the BBS and the BESTest. Test-retest reliability of the BESTest was calculated using ICC score, with a 95% CI. For interpretation of the ICC score, ranging from 0.00 to 1.00, values greater than 0.75 are considered good and values of 0.75 or less are considered
poor-to-moderate reliability.\textsuperscript{28} Construct validity was assessed by calculating the Pearson correlation coefficient ($r$) of the total scores of the BESTest and the BBS at baseline. We also calculated the Spearman rho ($r_s$) for the total score of the BESTest and the ordinal scores of the EDSS. We calculated the standard error of the measurement (SEM) based on the test-retest reliability, in which the “SEM is indicative of the range of scores that can be expected on retesting.”\textsuperscript{28(p609)} The SEM was calculated as follows:

$$SEM = SD \times \sqrt{(1 – ICC)}$$\textsuperscript{29}

The calculated SEM was used to determine the MDC at the 95\% and 90\% CIs as follows\textsuperscript{28}:

$$MDC = z \times SEM \times \sqrt{2}$$

$$MDC_{90} = 1.65 \times SEM \times \sqrt{2}$$

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

A receiver operating characteristic (ROC) curve was used to determine accuracy, or the ability of the BESTest to discriminate between fallers and nonfallers. Participants who indicated that they had fallen at least once in the previous month were considered fallers. The accuracy of the test was assessed using the area under the curve (AUC), which can be interpreted as the probability of correctly identifying participants with and without a history of falls. An AUC value of 0.9 to 1.0 indicates high accuracy, 0.7 to less than 0.9 indicates moderate accuracy, 0.5 to less than 0.7 indicates low accuracy, and less than 0.5 indicates that the test provides no better information than chance. The BESTest cutoff point used to classify a participant as a faller was chosen by selecting the score that provided the best sensitivity and specificity.\textsuperscript{28}

**Results**

Demographic characteristics and balance measure performance results for all the participants are summarized in Table 1. A subset of 12 individuals participated in the test-retest reliability testing (Table 2). The intrarater test-retest reliability of the BESTest was excellent: ICC\textsubscript{3,1} = 0.98 (95\% CI, 0.91-0.99). The construct validity testing of the BESTest with the BBS was completed for all 20 participants using the BBS as the reference standard. The BBS score was skewed (skewness = –1.7), which violated the assumption of normality. Therefore, the Spearman rho statistic was calculated. The BESTest score was highly correlated with the BBS score ($r = 0.94$, $P = .01$). Figure 1 provides a plot of the BESTest score and the BBS score by participant. The BESTest score was also strongly negatively correlated with the EDSS score ($r_s = –0.88$, $P = .01$), indicating that participants who have higher BESTest scores have lower EDSS scores.

<table>
<thead>
<tr>
<th>Participant No./Sex</th>
<th>BBS score</th>
<th>No. of falls reported</th>
<th>EDSS score</th>
<th>Initial BESTest score, %</th>
<th>BESTest retest score, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/F</td>
<td>53</td>
<td>2</td>
<td>6.0</td>
<td>78.7</td>
<td>83.3</td>
</tr>
<tr>
<td>2/F</td>
<td>54</td>
<td>0</td>
<td>6.0</td>
<td>82.4</td>
<td>88.0</td>
</tr>
<tr>
<td>3/F</td>
<td>56</td>
<td>0</td>
<td>3.0</td>
<td>96.3</td>
<td>98.2</td>
</tr>
<tr>
<td>4/M</td>
<td>56</td>
<td>2</td>
<td>6.0</td>
<td>78.8</td>
<td>83.3</td>
</tr>
<tr>
<td>6/F</td>
<td>46</td>
<td>0</td>
<td>6.5</td>
<td>63.9</td>
<td>52.8</td>
</tr>
<tr>
<td>7/M</td>
<td>29</td>
<td>2</td>
<td>6.5</td>
<td>31.5</td>
<td>38.0</td>
</tr>
<tr>
<td>9/F</td>
<td>34</td>
<td>4</td>
<td>6.5</td>
<td>41.5</td>
<td>43.5</td>
</tr>
<tr>
<td>10/F</td>
<td>49</td>
<td>1</td>
<td>5.0</td>
<td>75.0</td>
<td>75.0</td>
</tr>
<tr>
<td>13/F</td>
<td>15</td>
<td>0</td>
<td>6.5</td>
<td>39.0</td>
<td>44.4</td>
</tr>
<tr>
<td>14/F</td>
<td>56</td>
<td>0</td>
<td>2.5</td>
<td>97.2</td>
<td>93.5</td>
</tr>
<tr>
<td>19/F</td>
<td>56</td>
<td>0</td>
<td>2.0</td>
<td>83.3</td>
<td>85.2</td>
</tr>
<tr>
<td>20/F</td>
<td>54</td>
<td>4</td>
<td>4.0</td>
<td>83.3</td>
<td>83.3</td>
</tr>
</tbody>
</table>

Mean ± SD: 46 ± 13.9, 1.25 ± 1.5, 5.0 ± 1.7, 71.3 ± 22.1, 72.4 ± 21.5

Abbreviations: BBS, Berg Balance Scale; BESTest, Balance Evaluation Systems Test; EDSS, Expanded Disability Status Scale.

**Table 2. BESTest test-retest participant data**

**Figure 1. Plot of Berg Balance Scale and Balance Evaluation Systems Test (BESTest) scores by participant**

Red dots indicate individuals who reported falling within past month.
Reliability and Validity of BESTest in MS

No floor effect was noted for either test. The BBS scores were significantly skewed. Seven participants (35%) received the maximal score on the BBS, whereas none received the maximal score on the BESTest. The BBS demonstrated limited ability to classify fallers using the industry standard cutoff score of less than 45 of 56,0-32; the BBS correctly identified only 33.3% of participants with a fall history.

The ROC plot for the BESTest is shown in Figure 2. The BESTest demonstrated moderate accuracy in identifying individuals with a history of falls within the past month (AUC = 0.76). Nine participants (45%) had fallen in the month before the assessment, with 89% of those individuals reporting more than one fall during that time. We determined that a BESTest score of 81% can differentiate between fallers and nonfallers with 0.89 sensitivity and 0.64 specificity when used to assess functional balance in persons with MS and moderate disability. The SEM for the BESTest in this study was 4.16 points, which translates into an MDC_90 of 9.7% and an MDC_95 of 11.5%.

Discussion

This study adds to the emerging body of research supporting the use of the BESTest to reliably assess and capture true changes in functional balance in persons with MS. These results confirmed that the BESTest has excellent test-retest reliability (ICC_3,1 = 0.98) and strong construct validity (r = 0.88) in individuals with moderate MS-related disability. Previous studies examining the psychometric properties of the BESTest in persons with MS used similar sample sizes of individuals but with mild MS-related symptoms,18,20. The present results support recent work by Potter et al.,20 who reported that the BESTest has strong test-retest reliability (0.94) when used to evaluate functional balance in 21 ambulatory persons with mild MS-related disability. These findings are in agreement with those of previous studies in individuals with Parkinson’s disease, stroke, and other conditions that result in balance impairments.12,16,18

The present findings also support the construct validity of the tool as a balance measure for this population. In their preliminary study, Jacobs and Kasser18 reported that total BESTest scores significantly correlated (Pearson r = 0.48-0.76, P < .01) with the measures of postural performance (step initiation, limits of stability, and anticipatory postural adjustments) and EDSS scores (Spearman rho = 0.85, P < .0005) in a population of individuals with mild-to-moderate MS (median EDSS score = 2.5). Their average total BESTest score was lower for the persons with MS, and the BESTest score inversely correlated with disease severity as measured by the EDSS. The present findings also support a strong inverse relationship (r = –0.88) between the two measures.18 The average EDSS score for their study was 2.5 (range, 0-4.5), more than 2 points lower than the present participant average of 4.8 (range, 2.0-6.5), indicating that the BESTest is a valid balance assessment for persons with greater disease severity. Given the chronic nature of MS, this is an important finding.

Given the functional fluctuations associated with MS, evidence that the BESTest has strong test-retest reliability and a small MDC in this patient population is an important finding and provides clinicians with a measure that can assist in documenting real changes in balance performance. Using an ROC plot, we determined that a BESTest score of less than 81% correctly identified eight of nine (89%) fallers with moderate MS (mean EDSS score = 4.8). In the present sample, the BESTest performed better than the BBS, which correctly identified only 33.3% of individuals with a history of falling at least once in the previous month.
These findings support the existence of a ceiling effect with the BBS for identifying fall risk in the MS population.\textsuperscript{10,27} Although neither the BBS nor the BESTest exhibited a floor effect in this study, seven participants (35%) scored a perfect score of 56 on the BBS, with one reporting a history of falling. The average BESTest score was 71\% (range, 32\%-97\%), and no participant attained a perfect score. These findings support previous work suggesting a risk of a ceiling effect with the BBS in higher-functioning patients and specifically with the MS population.\textsuperscript{3,16,26,27} The BESTest subcategories and the variety of individual items also allow the clinician to thoroughly assess the multiple systems contributing to postural control and sufficiently challenge the individual to capture functional balance consistent with the day-to-day activities of a younger population.\textsuperscript{13,16} One of the unique features of the BESTest is the inclusion of anticipatory and reactive postural responses. Given the impact that delayed somatosensory conduction\textsuperscript{2,3} has on the effectiveness of anticipatory and reactive responses in persons with MS, this information is critical to assess in this population. This information will allow clinicians to better educate their clients and design targeted interventions to address this specific balance impairment contributing to fall frequency in the MS population.

We also determined that a BESTest score change greater than 9.7\% (90\% CI) and 11.5\% (95\% CI) in patients with moderate MS disability represents true change in performance that exceeds the BESTest measurement error. The present findings are slightly higher than the MDC\textsubscript{95} of 10 points or 9\% of the total BESTest reported by Potter et al.\textsuperscript{20} These findings are useful for clinically evaluating patient improvement or decline and may prove to be more sensitive than the EDSS score to track change in patient function and the effectiveness of physical therapy interventions. Although not in the MS population, these MDCs are in line with recent work indicating that the responsiveness of the BESTest in individuals with subacute stroke is a 10\% change in total score after 14 days of physical therapy intervention.\textsuperscript{14}

With fatigue\textsuperscript{13} being a primary MS disease-related symptom, the ability of the participants to tolerate a challenging 45-minute gait-and-balance assessment was an a priori concern. The length of the test did not prove to be a limiter for any of the participants. All 20 participants were able to complete both the BESTest and the BBS with a 10-minute rest between the two tests. This is an important finding given that 60\% of the participants had an average EDSS score greater than 4.5, indicating that MS limited their ambulation distance and quality.

In terms of limitations, the present findings should be interpreted within the context of the study design. Although the sample size was small, the participants represented a range of functional performance, with total BESTest scores ranging from 32\% to 97\%. The feasibility of using the 36-item BESTest with persons from mild-to-severe balance impairments who often easily fatigue is an important finding. We acknowledge that external validity is limited by the design using a single rater with expertise in assessment of individuals with MS and the use of both balance measures. The risk of recall bias on the part of the single rater cannot be ruled out. Individuals unfamiliar with the BESTest can avail themselves of the free tutorials offered on the tool’s website before use in the clinic. Another limitation is the use of retrospective self-reported fall data. However, we limited the period to 1 month to increase the accuracy of recall and assess the current risk of falls given the high rate of falls in the MS population and the fluctuating nature of the disease.\textsuperscript{11,27,34} The UIC FFM\textsuperscript{22} has not been validated in this population but is the tool used in our clinic to assess fear of falling. The authors recognize this limitation and suggest use of the Fall Efficacy Scale-International for future studies.\textsuperscript{35} The absence of prospective fall data limits the present findings in a population of individuals whose function can vary based on the waxing-and-waning nature of MS symptoms. To categorize the BESTest as an effective fall risk screening tool across the spectrum of persons with MS, a study with a larger

**PRACTICE POINTS**

- The Balance Evaluation Systems Test (BESTest), a 36-item functional balance assessment, has no ceiling effect, excellent test-retest reliability, and strong construct validity when used to assess balance and fall risk in persons with MS.
- A change in BESTest score of at least 10\% may indicate a true change in balance performance in persons with MS.
- Despite its length (45 minutes to administer) and varying levels of difficulty, the BESTest is not too fatiguing to use in a rehabilitation or outpatient therapy setting as a comprehensive balance assessment for persons with moderate MS.
sample size, across EDSS scores, and with monitoring of prospective fall data is necessary.

In conclusion, the BESTest demonstrates excellent test-retest reliability and good diagnostic utility (accuracy with high sensitivity) as a measure of functional balance in persons with moderate disability due to MS. A BESTest score change of at least 10% exceeds measurement error and may indicate a true change in balance performance in persons with moderate MS. The BBS has limited capacity to accurately measure functional balance and identify fall risk in this patient population.

**Financial Disclosures:** The authors declare no conflicts of interest.

**References**