Comparison of Reliability, Validity, and Responsiveness of the Mini-BESTest and Berg Balance Scale in Patients With Balance Disorders

Marco Godi, Franco Franchignoni, Marco Caligari, Andrea Giordano, Anna Maria Turcato, Antonio Nardone

M. Godi, PT, MS, Posture and Movement Laboratory, Division of Physical Medicine and Rehabilitation, Salvatore Maugeri Foundation–IRCCS, Veruno, Italy.

F. Franchignoni, MD, Unit of Occupational Rehabilitation and Ergonomics, Salvatore Maugeri Foundation–IRCCS. Mailing address: Fondazione Salvatore Maugeri, Clinica del Lavoro e della Riabilitazione–IRCCS, Via Revislate 13, 1-28010, Veruno, Italy. Address all correspondence to Dr Franchignoni at: franco.franchignoni@fsm.it.

M. Caligari, PT, Posture and Movement Laboratory, Division of Physical Medicine and Rehabilitation, Salvatore Maugeri Foundation–IRCCS.

A. Giordano, PhD, Unit of Bioengineering, Salvatore Maugeri Foundation–IRCCS.

A.M. Turcato, PT, Posture and Movement Laboratory, Division of Physical Medicine and Rehabilitation, Salvatore Maugeri Foundation–IRCCS.

A. Nardone, MD, PhD, Division of Physical Medicine and Rehabilitation, Salvatore Maugeri Foundation–IRCCS, and Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy.

[Godi M, Franchignoni F, Caligari M, et al. Comparison of reliability, validity, and responsiveness of the Mini-BESTest and Berg Balance Scale in patients with balance disorders. *Phys Ther.* 2013;93:158–167.]

Published Ahead of Print: September 27, 2012
Accepted: September 17, 2012
Submitted: April 14, 2012

Background. Recently, a new tool for assessing dynamic balance impairments has been presented: the 14-item Mini-BESTest.

Objective. The aim of this study was to compare the psychometric performance of the Mini-BESTest and the Berg Balance Scale (BBS).

Design. A prospective, single-group, observational design was used in the study.

Methods. Ninety-three participants (mean age=66.2 years, SD=13.2; 53 women, 40 men) with balance deficits were recruited. Interrater (3 raters) and test-retest (1–3 days) reliability were calculated using intraclass correlation coefficients (ICCs). Responsiveness and minimal important change were assessed (after 10 sessions of physical therapy) using both distribution-based and anchor-based methods (external criterion: the 15-point Global Rating of Change [GRC] scale).

Results. At baseline, neither floor effects nor ceiling effects were found in either the Mini-BESTest or the BBS. After treatment, the maximum score was found in 12 participants (12.9%) with BBS and in 2 participants (2.1%) with Mini-BESTest. Test-retest reliability for total scores was significantly higher for the Mini-BESTest (ICC=.96) than for the BBS (ICC=.92), whereas interrater reliability was similar (ICC=.98 versus .97, respectively). The standard error of measurement (SEM) was 1.26 and the minimum detectable change at the 95% confidence level (MDC_{95}) was 3.5 points for Mini-BESTest, whereas the SEM was 2.18 and the MDC_{95} was 6.2 points for the BBS. In receiver operating characteristic curves, the area under the curve was 0.92 for the Mini-BESTest and 0.91 for the BBS. The best minimal important change (MIC) was 4 points for the Mini-BESTest and 7 points for the BBS. After treatment, 38 participants evaluated with the Mini-BESTest and only 23 participants evaluated with the BBS (out of the 40 participants who had a GRC score of ≥3.5) showed a score change equal to or greater than the MIC values.

Limitations. The consecutive sampling method drawn from a single rehabilitation facility and the intrinsic weakness of the GRC for calculating MIC values were limitations of the study.

Conclusions. The 2 scales behave similarly, but the Mini-BESTest appears to have a lower ceiling effect, slightly higher reliability levels, and greater accuracy in classifying individual patients who show significant improvement in balance function.

Body balance relies on feedback circuits fed by the input from different receptors, including somatosensory, labyrinthine, and visual.[1] These inputs have to be adequately integrated in the central nervous system in order to produce appropriate changes in motor output to correct internal and external balance perturbations.[2] If one or more of these inputs, their integration, or the motor output are impaired, balance disorders occur.[3]

Because balance control is a complex task, simple tests of postural stability, such as one-leg stance, are not appropriate for a comprehensive assessment of patients with balance impairment.[4] People with balance disorders may be unstable in many different daily life situations (eg, when walking, when turning, when reaching for a far object, after an external perturbation).[5–7] Clinical scales have been developed to provide a comprehensive view of balance performances, as close as possible to real-life situations.[8] To evaluate postural stability in a more functional context, these clinical scales would appear to be more appropriate than simple tests of postural stability.

The Berg Balance Scale (BBS)[9] is one of the most widely used tools for balance assessment.[10] Its psychometric properties have been well assessed, and the scale has shown to be a valid and reliable measure of balance.[11] However, some important limitations of the BBS have been described, such as the need for some rescoring of the rating scale,[12] a ceiling effect,[11] and relatively low responsiveness.[13] Moreover, dynamic balance (eg, reacting to a perturbation, gait) is unexplored by the BBS.

Recently, a new clinical tool for assessing balance impairments has been presented: the Balance Evaluation Systems Test (BESTest).[14] This 36-item test, at variance with the BBS, also scores dynamic balance and gait performance, and it has shown good reliability and validity for assessing balance in individuals with Parkinson disease (PD).[15] However, the drawbacks of the BESTest are that it takes about 45 minutes to administer and it comprises multiple dimensions.[16] Thus, with the aid of factor analysis and Rasch analysis, a short form of the BESTest with 14 items only, named the Mini-BESTest, was produced, with improved rating category, high reliability, and structural validity.[16] The Mini-BESTest includes important aspects of dynamic balance control, such as the capability to react to postural perturbations, to stand on a compliant or inclined surface, and to walk while performing a cognitive task. All of these features of balance control are known to be important in assessing balance disorders in different types of patients and reflect balance challenges during activities of daily living.[14,17] Recent articles have been published[18–20] in which some important psychometric characteristics of the Mini-BESTest (eg, responsiveness) compared favorably with those of the BBS in patients with PD.

The aim of this study was to perform a head-to-head comparison of the psychometric performance of the Mini-BESTest and the BBS in a convenience sample of patients with balance disorders of different origins. For this purpose, we estimated interrater and test-retest reliability, concurrent validity, sensitivity to change, and responsiveness of both scales.

Method

Participants

Ninety-nine patients (mean age=66.1 years, SD=13.1; 56 women, 43 men) consecutively admitted to our free-standing rehabilitation center (320 beds) for assessment and rehabilitation treatment were recruited, representing a convenience sample of inpatients with balance disorders. Patients were referred from surrounding acute care hospitals and general practitioners and were screened for rehabilitation potential. The inclusion criterion was the ability to fully participate in the study procedures (eg, absence of severe cognitive impairments, tolerance of balance and gait tasks without fatigue). Of the 99 patients recruited, 2 were unable to perform the assessment due to the severity of their illness, and 4 declined to participate. Thus, 93 patients (mean age=66.2 years, SD=13.2; 53 women, 40 men) took part in the study. The participants' diagnoses were as follows: 25 had PD, 25 had hemiparesis (9 right, 12 left), 6 had multiple sclerosis, 5 had vestibular disorders, 6 had neuromuscular diseases, 8 had hereditary ataxia, 8 had sensorimotor polyneuropathy, 4 had central nervous system neoplasm, and 6 had unspecific age-related balance disorders. Prior to taking part in the study, all participants signed an informed consent statement that had been approved by the Central Ethics Committee of the Salvatore Maugeri Foundation.

Assessment

Mini-BESTest. The Mini-BESTest is a 14-item balance scale that takes about 15 minutes to administer, is unidimensional, and is highly reliable.[16] It contains items covering a broad spectrum of performance tasks, including transitions and anticipatory postural adjustments, postural responses to perturbation, sensory orientation while standing on a compliant or inclined base of support, and dynamic stability in gait. Items are scored from 0 (unable to perform or requiring help) to 2 (normal performance). The maximum total score is 28.

BBS. The BBS is the most widely used and validated instrument for assessing balance performance in neurological conditions.[9] It is composed of 14 items that require subjects to maintain positions of varying difficulty and perform specific tasks such as standing and sitting unsupported, transfers (sit to stand and stand to sit), turn to look over shoulders, pick up an object from the floor, turn 360° and place alternate feet on a stool. Scoring is based on the subject's ability to perform the 14 tasks independently and/or meet certain time or distance requirements. Each item is scored on a 5-point ordinal scale ranging from 0 (unable to perform) to 4 (normal performance) so that the aggregate score ranges from 0 to 56.

Global Rating of Change (GRC). The GRC is a rating scale designed to quantify patients' improvement or deterioration over time. It is used to determine the effect of an intervention or chart the clinical course of a condition. The GRC was completed at the time of the final assessment (after the rehabilitation treatment) by each participant and the treating physical therapist. Participants were asked to independently rate the overall change in their balance from when they began treatment using a 15-point scale ranging from −7 ("a very great deal worse") to +7 ("a very great deal better"), with 0 indicating "unchanged."[21,22] We decided to use 2 external indicators (clinician and patient rating, respectively) because the use of independent anchors is recommended[23] and may reduce problems reported when using only the patient GRC.[21] Therefore, the mean value of the 2 GRC scores (physical therapist and patient) was used as a reference standard: participants with a rating from 0 to +3 ("a little bit better") were considered to have minimally changed or not changed, and those with a rating greater than 3 were considered moderately to largely improved.[24]

Procedure

All participants were evaluated with the Mini-BESTest and the BBS by the same rater before and after a physical therapy program for balance disorders. The raters for all procedures were 3 licensed physical therapists (M.G., M.C., and A.M.T.) who were specifically trained in administering the 2 balance scales. The raters were always blinded to their previous ratings.

For both the Mini-BESTest and the BBS, test-retest reliability and interrater reliability were analyzed in a subset of 32 consecutive participants (mean age=67.3 years, SD=13.5; 19 women, 13 men; 8 with PD, 7 with hemiparesis, 10 with other neurological disorders, 3 with vestibular disorders, and 4 with age-related balance disorders). For interrater reliability, each of the 3 physical therapists performed a simultaneous independent balance assessment at baseline; for test-retest reliability, participants were reassessed (by 1 of the 3 therapists) after 1 to 3 days. This sample size was determined on the basis of a pilot study, expecting to obtain intraclass correlation coefficient (ICC) values of about .90, with a 95% confidence interval (CI) of .20.[25]

The physical therapy program consisted of ten 1-hour sessions for 2 weeks of the following exercises: (1) static and dynamic functional balance activities (eg, reaching while standing, standing on one leg, sit-to-stand maneuver, turning, walking training); (2) exercises for training specific balance skills (eg, "push and release" techniques, stance on a foam surface, dual-task training); (3) flexibility and strength training; and (4) perturbation-based training on a platform continuously moving on the horizontal plane.[14,26,27] Each treatment session was individually tailored according to the participant's functional status and clinical indications.

At the end of the treatment, the GRC was completed by each participant and by the treating physical therapist (4 different physical therapists who were not involved in the study procedures). The participants and therapists were unaware of each other's responses.

Data Analysis

Descriptive statistics, including central tendency (median) and spread (25th–75th percentiles), were calculated for both balance scales and the GRC. Floor and ceiling effects were analyzed, calculating the percentages of individuals obtaining the lowest and the highest scores for the 2 scales. The Stata/IC version 10.1 software package (StataCorp LP, College Station, Texas) was used for the statistical analyses.

Reliability. The internal consistency of the Mini-BESTest and the BBS was assessed by means of the Cronbach alpha coefficient at both baseline and follow-up. Alpha values $\geq$.70 are recommended for group-level comparison, whereas a minimum of .85 to .90 is desirable for individual judgments.[28]

For both scales, test-retest and interrater reliability of global scores was calculated, using the ICC (2,1) and corresponding CI. For clinical measurements, ICC values should exceed .90 to ensure reasonable reliability.[29] Z-transformed ICCs obtained with 1,000 bootstrap samples were used to test ICC difference between measures.[30]

Validity. Convergent validity was assessed by calculating the Pearson correlation coefficient (r) of the total scores of the Mini-BESTest and the BBS (at both the first evaluation and

follow-up) and their changes (after versus before rehabilitation). Confidence intervals and comparisons of the correlation coefficients between the measures were calculated.[31]

In addition, because the GRC was considered the anchor (ie, the reference standard against which we judged whether a real improvement in the participants had occurred), it was used to provide a valid assessment of the same construct measured by the tools under longitudinal investigation.[24] Thus, a Pearson correlation between the GRC (mean value of the participant's and therapist's scores) and the change (after versus before rehabilitation) in the 2 balance scales was calculated and tested for differences between measures. Moreover, the correlation between the GRC rated by the participant and that rated by the physical therapist was used to investigate their relationship. For all of these correlations, we expected a "nontrivial" association between measures (ie, $r>.30$).[23]

Responsiveness. There are 2 types of approach for evaluating responsiveness and clinical significance[23]: distribution-based methods and anchor-based methods. The distribution-based methods are based on the statistical characteristics of the obtained sample and analyze the ability to detect change in general. The anchor-based methods require an external criterion to determine whether changes in outcome scores are clinically meaningful. We used both approaches in order to have a wide range of results on which to draw inferences about the minimal important change (MIC) for both scales, aware of the large variation and lack of convergence that these different methods could show.[32]

For the distribution-based methods, we calculated the standard error of measurement (SEM), which links the reliability of the measurement instrument to the standard deviation of the population.[33] The SEM and its CI were calculated on the basis of the analysis of variance used to produce the ICC.[34] Starting from the SEM, we calculated the minimum detectable change (MDC). The MDC represents the smallest change in score that likely reflects true change rather than measurement error alone. The calculation is the result of the multiplication of the SEM $\times$ z value $\times$ $\sqrt{2}$. The 95% confidence level (MDC_{95}) was established, corresponding to a z value of 1.96. As an example, if a participant has a change score equal to or above the MDC_{95} threshold, it is possible to state with 95% confidence that this change is reliable and not due to an error.

The second approach for evaluating responsiveness is the use of anchor-based methods. These methods were based on GRC assessment as an external criterion. The following 2 parameters were analyzed: (1) for the mean change approach, we calculated the mean change of participants graded on the GRC as not improved (GRC $\leq$3), moderately improved (3<GRC<5), or largely improved (GRC $\geq$5); and (2) for the receiver operating characteristic (ROC) curve approach,[29] we determined the optimal cutoff score and the area under the curve (AUC) after having split the participants based on a GRC $\leq$3 or higher, and thus having considered a GRC >3 as an index of meaningful change.

A ROC curve plots sensitivity (y-axis) against 1 $-$ specificity (x-axis). In this context, sensitivity was calculated as the number of participants correctly identified as improved based on the cutoff value divided by all participants identified as having undergone a meaningful change (GRC >3), whereas specificity refers to the number of participants who were correctly identified as not improved based on the cutoff value divided by all participants who truly did not undergo a meaningful change (GRC $\leq$3). The optimal cutoff score was chosen as the point that jointly maximized sensitivity and specificity (being associated with the least amount of misclassification).

The AUC can be interpreted as the probability of correctly identifying a patient who has improved in randomly selected pairs of patients who have and have not shown an improvement. The greater the AUC, the greater a measure's ability to distinguish patients who improved from those who do not improve; as a general rule, an AUC >0.8 is considered to have excellent discrimination.[29] Based on the study by Turner et al,[24] our ROC analysis used the entire cohort in order to increase precision and obtain more logical estimates of the MIC values.

Formal testing for a difference in the AUCs between scales was performed according to the procedure of DeLong et al.[35] To obtain CIs for the ROC analysis results, we drew 500 bootstrap samples and calculated the AUC, as well as the sensitivity and specificity values associated with the best cutoff scores in each bootstrap replication. The mean of the bootstrap values was taken as the best estimate, with the CI calculated as 1.96 $\times$ SD (as an estimate of the standard error) of the 500 bootstrap values.[32]

Role of the Funding Source

This study was supported, in part, by "Giovani Ricercatori 2009" grant (GR-2009-1471033) to Mr Godi and by "Progetto Strategico 2007" grant (RFPS-2007-1-641398) to Dr Nardone from the Italian Ministry of Health. The study sponsor was not involved in: study design; collection, analysis,

Table 1.
Descriptive Statistics Related to Values of the Mini-BESTest, the Berg Balance Scale (BBS), and the Global Rating of Change (GRC) in the Whole Group (n=93) and to Values of the Mini-BESTest and the BBS in the Test-Retest and Interrater Reliability Subgroup (n=32)

Measure	Minimum	Maximum[a]	$\bar{X}$	SD	1st Quartile	Median	3rd Quartile
Mini-BESTest							
Baseline	1	27	12.8	6.9	8	12	19
After treatment	1	28 (2)	15.8	6.9	11	15	22
Change	−1	10	3.1	2.4	1	4	5
Test-retest and interrater reliability subgroup	1	25	11.1	7.6	5	11	15
BBS							
Baseline	4	55	42	11.2	38	45	50
After treatment	4	56 (12)	46.3	10.3	42	49	54
Change	−2	17	4.2	3.9	1	4	6
Test-retest and interrater reliability subgroup	4	55	38.4	14.2	30	42	48
GRC	0	6	2.9	1.2	2	3	3.5

[a] Number of participants recording a ceiling effect shown in parentheses.

or interpretation of data; writing of the report; or the decision to submit the manuscript for publication.

Results

Descriptive Statistics

Table 1 provides the descriptive statistics for 3 measures (both at baseline and after treatment for the Mini-BESTest and the BBS and only after treatment for the GRC) in the whole group (n=93) and for Mini-BESTest and the BBS in the test-retest and interrater reliability subgroup (n=32). No clinical problems were encountered during assessment procedures. No dropouts occurred.

Figure 1 shows the score distribution of the 2 scales before and after treatment. In both the Mini-BESTest and the BBS, neither top scores at baseline nor floor scores at any time were found. After treatment, 12 participants (12.9%) reached the maximum BBS score, whereas 2 participants (2.1%) reached the Mini-BESTest top score (Tab. 1).

The mean GRC was ≤3 in 53 participants (57%, small or null improvement), 3<GRC<5 in 34 participants (36.5%, moderate improvement), and GRC ≥5 in 6 participants (6.4%, large improvement). No participants worsened according to the GRC.

Reliability

There was a statistically significant difference in test-retest reliability between the Mini-BESTest and the BBS, whereas both Cronbach alpha and interrater reliability were similar in both groups (Tab. 2).

Validity

The scores of the Mini-BESTest and the BBS were highly correlated at both baseline and follow-up (for both, r=.85, CI=.78–.90) (Fig. 2). The correlation between score changes of the Mini-BESTest and the BBS over the course of the rehabilitation program was r=.58 (P<.001).

The correlation between mean GRC and the score changes (after versus before rehabilitation) was r=.72 (CI=.61–.81) for the Mini-BESTest and r=.62 (CI=.48–.73) for BBS; the difference between the correlation coefficients was not statistically significant. The GRC rated by the participant and that by the physical therapist were significantly correlated (r=.61, P<.001).

Responsiveness

Distribution-based methods. The SEM and MDC_{95} values for both the Mini-BESTest and the BBS are shown in Table 2.

Anchor-based methods. For both scales, the mean score changes in those participants who were rated as having a small or null improvement (GRC ≤3), moderate improvement (3<GRC<5), or large improvement (GRC ≥5) are shown in Table 2.

Splitting data according to the presence of a moderate to large GRC improvement (GRC ≤3 versus GRC >3), both AUCs were high and similar (Tab. 2, Fig. 3). The cutoff score that best identified meaningful improvement in clinical status (as measured by GRC >3) was 4 points for the Mini-BESTest and 6 points for the BBS.

Overall, a MIC value of 4 points for the Mini-BESTest and 7 points for the BBS represented the best triangulation of these results, adopting values

higher than the respective MDC_{95} value for each scale. Among the 40 participants who had a moderate to large improvement in balance (GRC >3) after physical therapy, 38 showed a change of ≥4 points on the Mini-BESTest, whereas only 23 showed a change of ≥7 points on the BBS.

Discussion

Valid inferences about the efficacy of treatment trials require high-quality outcome measures that meet rigorous measurement standards. The present study was conducted to analyze reliability and validity issues in both the Mini-BESTest and the BBS and to compare their responsiveness after a 10-session physical therapy program for balance disorders. Our results are in line with the recent literature[13,18,19] and indicate that the Mini-BESTest shows sound psychometric properties, which compare favorably with those of the BBS, particularly when measuring change at the individual level.

At the follow-up evaluation, 2 participants (2.1%) reached the top score on the Mini-BESTest, whereas 12 participants (about 13%) reached the maximum score on the BBS. Our findings are in agreement with those of previous studies that showed the BBS to have a ceiling effect in people with PD, as well as in other populations.[13,18,19] Recently, in people with PD, a lesser ceiling effect and skewed distribution were found for the Mini-BESTest with respect to the BBS.[13] Usually only subgroups of patients with severely limited function do not show a ceiling effect on the BBS.[18] This fact raises an important concern about the use of the BBS as an outcome measure to evaluate balance impairments: it represents a limited ability of the tool to discriminate among patients with quite good balance function. On the contrary, the absence of a significant ceiling compression effect on the Mini-BESTest speaks in favor of the use of this scale, which represents a more comprehensive measure of balance, with items (eg, compensatory steps, walking with dual task) that are able to challenge patients with even minimal impairment in balance function.

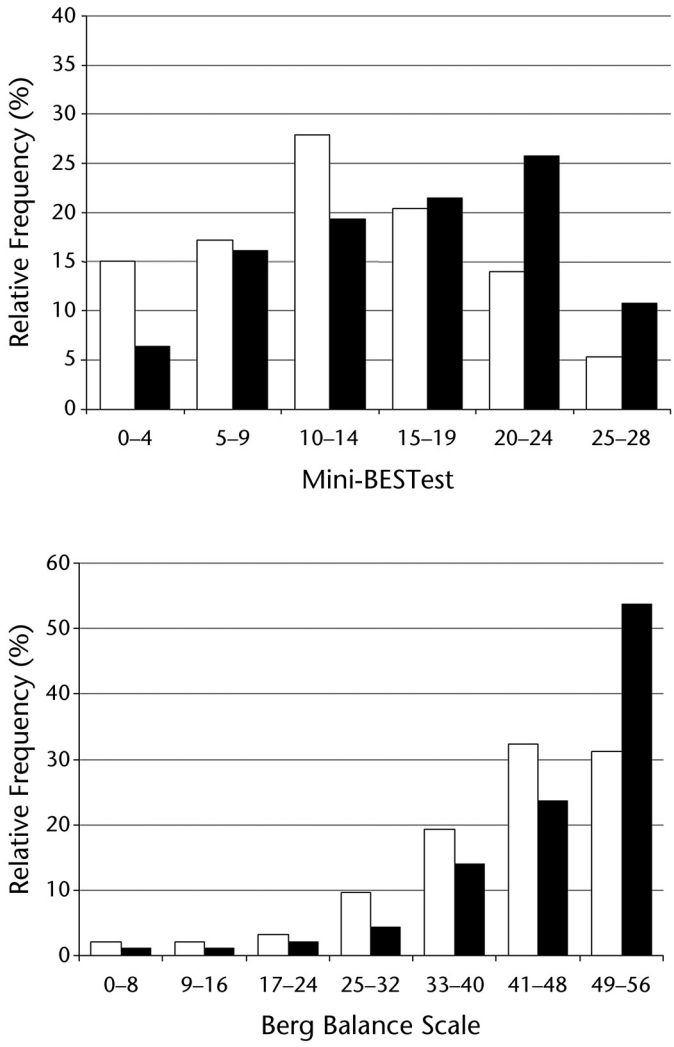

Figure 1.
Histogram of grouped frequency distribution (%) for Mini-BESTest scores (range=0–28) and Berg Balance Scale scores (range=0–56), before (white columns) and after (black columns) physical therapy program.

Reliability

The Cronbach alpha showed high values (≥.90) in both tests. On the basis of the ICC, both the Mini-BESTest and the BBS performed very well in terms of test-retest reliability (.96 versus .92) and interrater reliability (.98 versus .97).

The high reliability of both balance scales is in accordance with previous findings. A recent study[19] performed in individuals with PD demonstrated similar levels of reliability for the Mini-BESTest (interrater reliability ICC=.91, test-retest reliability ICC=.92). In earlier reliability studies using the BBS, test-retest ICCs ranged from .80 to .99,[11,15,36] whereas interrater ICCs were usually ≥.95.[36,37]

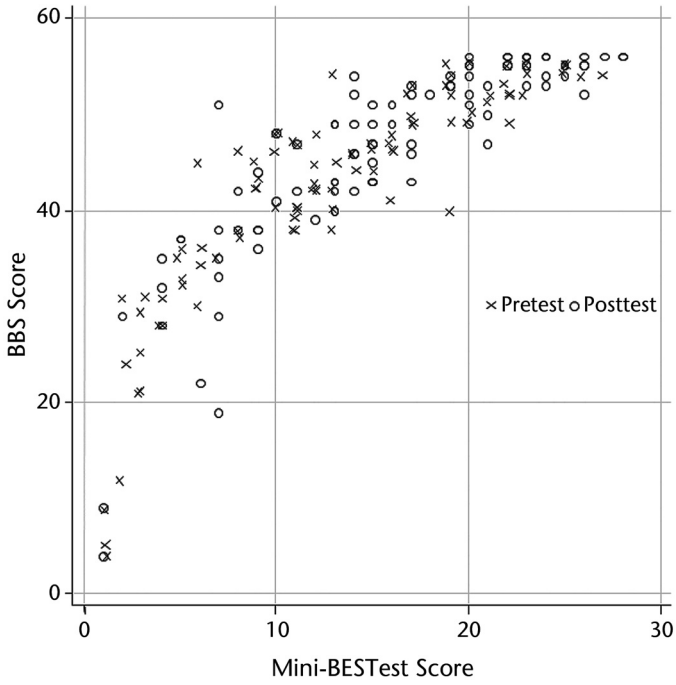

Figure 2.
Scatterplot showing the relationship between the Mini-BESTest and the Berg Balance Scale (BBS) raw scores, before and after the physical therapy program.

Validity

The high correlation between the 2 scales supports the convergent validity of the MiniBESTest with the BBS, the most commonly used scale for balance assessment. A high correlation between the 2 scales also was found in a recent study of individuals with PD,[18] and a high correlation of the Mini-BESTest with the BBS, the Timed "Up & Go" Test, and the Falls Efficacy Scale was reported in individuals with both PD and stroke.[38] In addition, the ability of both the participants and the physical therapist to acceptably estimate the change in balance performance (during a 2-week transition period) is confirmed by the correlation of their GRC assessments with each other and with change in the Mini-BESTest and BBS scores.

Responsiveness

If rating scales are used as primary outcome measures in clinical studies, there is a need to know the extent to which changes in their scores reflect clinically important changes in patients' health status. There is a lack of consensus regarding the best method to determine the MIC, and a recent study recommended using multiple approaches followed by a triangulation to obtain one value or a small range of values for the MIC,[32] as we did in the present study.

Distribution-based methods. The MDC_{95} value was 3.5 points for the Mini-BESTest and 6.2 points for the BBS. In the only study that had sufficient data to calculate the MDC for the Mini-BESTest,[19] this value was about 4 (ie, very close to our result). Romero et al[39] recently found an MDC_{95} value of 6.5 points for the BBS and noted that this value was not constant across different levels of function, being lower in individuals with better performance. Our findings also appear to be confirmed by the observation that reported MDC_{90-95} values for the BBS range from 5 to 8 points.[36,39–41]

Anchor-based methods. The mean score change in participants who were rated as having had a moderate improvement ($3<GRC<5$) was 4.6 points for the Mini-BESTest and 7.0 points for the BBS. Using ROC curves, the relative discriminatory accuracy of the 2 tests was excellent (>90%) and statistically equivalent. The Mini-BESTest showed a higher sensitivity than the BBS (94% versus 77%, respectively) (Tab. 2), which indicates a higher capacity to identify those participants who underwent a clinically important change, which is crucial in clinical settings. Likewise, Duncan et al[20] found a comparable accuracy of the 2 tests in predicting individuals with PD who were prone to falling at 6 months, whereas King et al[18] reported that the Mini-BESTest was slightly more successful than the BBS at discriminating subgroups of PD severity as measured by the Hoehn and Yahr scale.

In general, the results of anchor-based methods (and related values of MIC) should be considered more important than those of the distribution-based methods (including values of MDC),[23] and—as Turner et al[24] stated—distribution-based approaches should act only as a temporary surrogate, pending availability of empirically established anchor-based MIC values. However, the large variations of MIC indexes that can be found among populations and methods[32] indicate that in the puzzle to establish the MIC, we should select only MIC values that are above the MDC.[24]

Accordingly, the overall results of our study suggest a change of 4 points in the Mini-BESTest as the most acceptable MIC value. The MIC value was higher than MDC_{95} value for this scale and represents a score

change just slightly lower than the mean change in our group of participants who showed a moderate balance improvement (corresponding to 3<GRC<5). Similarly, in our sample, a change of 7 points appears to be the most adequate MIC value for the BBS: again, it was higher than its MDC_{95} value (6.2 points) and corresponds to the mean change in our participants who showed a moderate balance improvement. Furthermore, these MIC values represent a change of similar size on the 2 scales. A change of 4 points represents a variation of about 14% for the Mini-BESTest (maximum score: 28), and a change of 7 points represents a variation of 13% for the BBS (maximum score: 56). However, switching from group level to person level, 38 (95%) of the 40 participants who had a moderate to large improvement in balance (GRC ≥3.5) showed a change after physical therapy equal to or higher than the MIC value (4 points) for the Mini-BESTest, whereas only 23 (58%) showed a change equal to or higher than the MIC value (7 points) for the BBS.

These findings are the first analyzing in depth the responsiveness of the Mini-BESTest and are in line with those concerning the BBS. Moreover, Romero et al[39] recently underscored that measurement error (and parameters derived from it) often are not constant across different levels of function and related scores. As a consequence, caution is mandatory when interpreting and using these MIC values in different populations and settings, particularly considering the intrinsic weaknesses of GRC.[21,42] The GRC (and the MIC values derived from it) suffers from the problem of the subjective retrospective judgments of change (eg, due to "recall bias," or problematic patient ability to understand the context of improvement).[21] To reduce these drawbacks, we used the mean of 2 ratings (participant and therapist),

Table 2.
Reliability and Responsiveness Indexes for Mini-BESTest and Berg Balance Scale[a]

Variable	Mini-BESTest	BBS
Reliability		
Cronbach alpha: baseline/follow-up	.90/.91	.93/.93
Test-retest reliability: ICC	*.96 (.94–.99)*[b]	*.92 (.87–.97)*[b]
Interrater reliability: ICC	.98 (.97–.99)	.97 (.96–.99)
Responsiveness: distribution-based methods		
SEM	1.26 (1.01–1.65)	2.18 (1.76–2.87)
MDC_{95}	3.5	6.2
Responsiveness: anchor-based methods		
Mean score change in patients with:		
• null/small improvement (GRC ≤3)	1.6	1.9
• moderate/medium improvement (3<GRC<5)	4.6	7.0
• large improvement (GRC ≥5)	7.0	9.2
Area under the ROC curve	0.92 (0.84–0.97)	0.91 (0.84–0.98)
Sensitivity	94 (87–100)	77 (65–89)
Specificity	81 (70–92)	97 (92–100)
Optimal cutoff score	4 (3.0–4.9)	6 (4.4–7.6)

[a] Data were calculated on the whole sample (n=93), except for test-retest and interrater reliability (n=32); 95% confidence intervals are shown in parentheses. ICC=intraclass correlation coefficient, SEM=standard error of measurement, MDC_{95}=minimum detectable change at 95% confidence interval, GRC=Global Rating of Change, ROC=receiver operating characteristic.
[b] Italics denote significant difference between measures (*P*<.001).

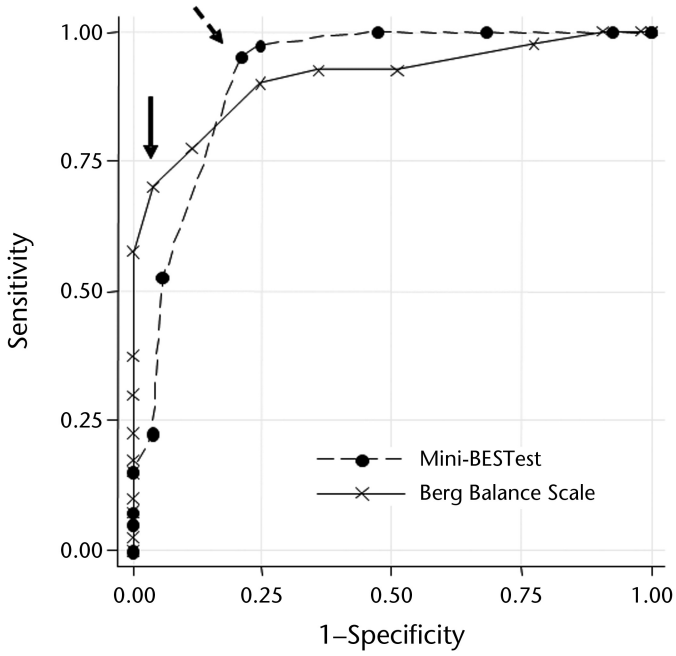

Figure 3.
Comparison between the receiver operating characteristic curves of the Mini-BESTest and the Berg Balance Scale, showing their overall accuracy in identifying a balance improvement according to a Global Rating of Change score of ≤3 versus >3. Arrows show the point that jointly maximizes sensitivity and specificity.

after reporting the correlations between them.

An additional limitation of the present study is the selection criteria of our convenience sample (recruited with a consecutive sampling method), which may represent a threat to external validity. Our sample was a cross-section of adults drawn from a single rehabilitation facility and with balance disorders of very different origins and severities. Finally, even if raters were blinded to their previous ratings, a memory effect cannot be ruled out.

In conclusion, this study showed—within the context analyzed and our specific patient group—the high reliability levels of the Mini-BESTest, confirmed those of the BBS, and proved the validity of both scales for measuring balance function and its change over time. In addition, our findings show how much the calculation of success rates (ie, percentages of patients having a change greater than the MIC value) can be useful from a clinical point of view.[32]

Most responsiveness indexes of the Mini-BESTest were equivalent or compared favorably with those of the BBS. The main advantages of the Mini-BESTest over the BBS appear to be that it has a lower ceiling effect together with slightly higher reliability levels, which led to greater accuracy in classifying individual patients who showed significant improvement in balance function.

Further studies are needed to confirm and expand the present results (to increase their generalizability), including analyses based on Rasch-transformed rating scores. Nevertheless, our results for the Mini-BESTest are in line with those of previous studies conducted in different countries and contexts using the same instrument, thus increasing our confidence in the relative validity of these findings.

Mr Godi, Dr Franchignoni, Mr Caligari, and Dr Nardone provided concept/idea/research design. Mr Godi, Dr Franchignoni, Mr Caligari, Dr Giordano, and Dr Nardone provided writing and data analysis. Mr Godi, Mr Caligari, and Ms Turcato provided data collection. Dr Franchignoni and Dr Nardone provided project management and study participants. Dr Nardone provided facilities/equipment and institutional liaisons. Ms Turcato and Dr Nardone provided consultation (including review of manuscript before submission).

This work was supported, in part, by "Giovani Ricercatori 2009" and "Progetto Strategico 2007" grants from the Italian Ministry of Health.

DOI: 10.2522/ptj.20120171

References

1 Johansson R, Magnusson M. Human postural dynamics. *Crit Rev Biomed Eng*. 1991;18:413–437.

2 Goodworth AD, Peterka RJ. Sensorimotor integration for multi-segmental frontal plane balance control in humans. *J Neurophysiol*. 2012;107:12–28.

3 Buchanan JJ, Horak FB. Voluntary control of postural equilibrium patterns. *Behav Brain Res*. 2003;143:121–140.

4 Briggs RC, Gossman MR, Birch R, et al. Balance performance among noninstitutionalized elderly women. *Phys Ther*. 1989;69:748–756.

5 Czernuszenko A, Członkowska A. Risk factors for falls in stroke patients during inpatient rehabilitation. *Clin Rehabil*. 2009;23:176–188.

6 Orr R. Contribution of muscle weakness to postural instability in the elderly: a systematic review. *Eur J Phys Rehabil Med*. 2010;46:183–220.

7 Plotnik M, Giladi N, Dagan Y, Hausdorff JM. Postural instability and fall risk in Parkinson's disease: impaired dual tasking, pacing, and bilateral coordination of gait during the "On" medication state. *Exp Brain Res*. 2011;210:529–538.

8 Yelnik A, Bonan I. Clinical tools for assessing balance disorders. *Neurophysiol Clin*. 2008;38:439–445.

9 Berg KO, Wood-Dauphinée SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. *Can J Public Health*. 1992;83(suppl 2):S7–S11.

10 Tyson SF, Connell LA. How to measure balance in clinical practice: a systematic review of the psychometrics and clinical utility of measures of balance activity for neurological conditions. *Clin Rehabil*. 2009;23:824–840.

11 Blum L, Korner-Bitensky N. Usefulness of the Berg Balance Scale in stroke rehabilitation: a systematic review. *Phys Ther*. 2008;88:559–566.

12 Kornetti DL, Fritz SL, Chiu YP, et al. Rating scale analysis of the Berg Balance Scale. *Arch Phys Med Rehabil*. 2004;85:1128–1135.

13 Pardasaney PK, Latham NK, Jette AM, et al. Sensitivity to change and responsiveness of four balance measures for community-dwelling older adults. *Phys Ther*. 2012;92:388–397.

14 Horak FB, Wrisley DM, Frank J. The Balance Evaluation Systems Test (BESTest) to differentiate balance deficits. *Phys Ther*. 2009;89:484–498.

15 Leddy AL, Crowner BE, Earhart GM. Functional gait assessment and balance evaluation system test: reliability, validity, sensitivity, and specificity for identifying individuals with Parkinson disease who fall. *Phys Ther*. 2011;91:102–113.

16 Franchignoni F, Horak F, Godi M, et al. Using psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. *J Rehabil Med*. 2010;42:323–331.

17 Horak FB. Postural orientation and equilibrium: what do we need to know about neural control of balance to prevent falls? *Age Ageing*. 2006;35:ii7–ii11.

18 King LA, Priest KC, Salarian A, et al. Comparing the Mini-BESTest with the Berg Balance Scale to evaluate balance disorders in Parkinson's disease. *Parkinsons Dis*. 2012;2012:375419. Epub 2011 Oct 24.

19 Leddy AL, Crowner BE, Earhart GM. Utility of the Mini-BESTest, BESTest, and BESTest sections for balance assessments in individuals with Parkinson disease. *J Neurol Phys Ther*. 2011;35:90–97.

20 Duncan RP, Leddy AL, Cavanaugh JT, et al. Accuracy of fall prediction in Parkinson disease: six-month and 12-month prospective analyses. *Parkinsons Dis*. 2012;2012:237673. Epub 2011 Nov 30.

21 Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther*. 2009;17:163–170.

22 Jaeschke R, Singer J, Guyatt G. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407–415.

23 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61:102–109.

24 Turner D, Schünemann HJ, Griffith LE, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *J Clin Epidemiol*. 2009;62:374–379.

25 Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*. 2002;21:1331–1335.

26 Shubert TE. Evidence-based exercise prescription for balance and falls prevention: a current review of the literature. *J Geriatr Phys Ther*. 2011;34:100–108.

27 Corna S, Nardone A, Prestinari A, et al. Comparison of Cawthorne-Cooksey exercises and sinusoidal support surface translations to improve balance in patients with unilateral vestibular deficit. *Arch Phys Med Rehabil*. 2003;84:1173–1184.

28 Bland JM, Altman DG. Cronbach's alpha. *BMJ*. 1997;314:572.

29 Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Prentice Hall Health; 2009.

30 Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. *Spine*. 2000;25:2095–2102.

31 Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 3rd ed. Shelton, CT: PMPH USA Inc; 2008.

32 Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol*. 2010;63:524–534.

33 de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes*. 2006;4:54.

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

35 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.

36 Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-Item Short-Form Health Survey, and the Unified Parkinson Disease Rating Scale in people with parkinsonism. *Phys Ther*. 2008;88:733–746.

37 de Figueiredo KM, de Lima KC, Cavalcanti Maciel AC, Guerra RO. Interobserver reproducibility of the Berg Balance Scale by novice and experienced physiotherapists. *Physiother Theory Pract*. 2009;25:30–36.

38 Bergström M, Lenholm E, Franzén E. Translation and validation of the Swedish version of the mini-BESTest in subjects with Parkinson's disease or stroke: a pilot study. *Physiother Theory Pract*. 2012;28:509–514.

39 Romero S, Bishop MD, Velozo CA, Light K. Minimum detectable change of the Berg Balance Scale and Dynamic Gait Index in older persons at risk for falling. *J Geriatr Phys Ther*. 2011;34:131–137.

40 Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. *Aust J Physiother*. 2001;47:29–38.

41 Conradsson M, Lundin-Olsson L, Lindelöf N, et al. Berg Balance Scale: intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. *Phys Ther*. 2007;87:1155–1163.

42 Cook CE. Clinimetrics corner. The Minimal Clinically Important Change Score (MCID): a necessary pretense. *J Man Manip Ther*. 2008;16:E82–E83.