EUROPEAN JOURNAL OF PHYSICAL AND REHABILITATION MEDICINE
EDIZIONI MINERVA MEDICA

Is the Berg Balance Scale an effective tool for the measurement of early postural control impairments in patients with Parkinson's Disease? Evidence from Rasch analysis

Fabio LA PORTA, Andrea GIORDANO, Serena CASELLI, Calogero FOTI, Franco FRANCHIGNONI

Eur J Phys Rehabil Med 2015 Sep 02 [Epub ahead of print]

EUROPEAN JOURNAL OF PHYSICAL AND REHABILITATION MEDICINE

Rivista di Medicina Fisica e Riabilitativa dopo Eventi Patologici
pISSN 1973-9087 - eISSN 1973-9095
Article type: Original Article

The online version of this article is located at http://www.minervamedica.it

Is the Berg Balance Scale an effective tool for the measurement of early postural control impairments in patients with Parkinson's Disease? Evidence from Rasch analysis

F. La Porta (MD)[1,2], A. Giordano (PhD)[3], S. Caselli (PT)[1],

C. Foti (MD)[1,4], F. Franchignoni (MD)[5]

[1] Rehabilitation Medicine Unit, Azienda USL Modena, Modena, Italy.

[2] PhD School in Advanced Sciences and Technologies in Rehabilitation Medicine and Sports, Tor Vergata University, Rome, Italy

[3] Unit of Bioengineering – Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno, Italy

[4] Chair of Rehabilitation Medicine, Tor Vergata University, Rome, Italy

[5] Unit of Occupational Rehabilitation and Ergonomics, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno, Italy,

Congresses: None

Funding: None

Conflicts of interest: None

Acknowledgements: Authors are grateful to dr. Rosemary Allpress for her assistance in the preparation of the manuscript.

Corresponding author:

F. La Porta

Rehabilitation Medicine Unit,

Azienda USL Modena,

Via Giardini 1355

41123

Modena

Italy

fabiolaporta@mail.com

ABSTRACT

Background. It is unclear whether the BBS is an effective tool for the measurement of early postural control impairments in patients with Parkinson's Disease (PD).

Aim. To evaluate BBS' content validity, internal construct validity, reliability and targeting in patients with PD within the Rasch analysis framework.

Design. Observational, cross-sectional study.

Setting. Outpatient Rehabilitation Unit.

Population. A sample of 285 outpatients with PD.

Methods. The content validity of the BBS was assessed using standard linking techniques. The BBS was administered by trained physiotherapists. The data collected then underwent Rasch analysis.

Results. Content validity analysis showed a lack of items assessing postural responses to tripping and slips and stability during walking. On Rasch analysis, the BBS failed the requirements of monotonicity, local independence, unidimensionality and invariance. After rescoring 7 items, grouping of locally dependent items into testlets, and deletion of the static sitting balance item because mistargeted and underdiscriminating, the Rasch-modified BBS for PD (BBS-PD) showed adequate internal construct validity (χ^2_{24}=39.693; p=0.023), including absence of differential item functioning (DIF) across gender and age, and was, as a whole, sufficiently precise for individual person measurement (PSI=0.894). However, the scale was not well targeted to the sample in view of the prevalence of higher scores.

Conclusion. This study demonstrated the internal construct validity and reliability of the

BBS-PD as a measurement tool for patients with PD within the Rasch analysis framework. However, the lack of items critical to the assessment of postural control impairments typical of PD, affected negatively the targeting, so that a significant percentage of patients was located in the higher ability range of the measurement continuum, where precision of measurement is reduced.

Clinical rehabilitation impact. These findings suggest that the BBS, even if modified, may not be an effective tool for the measurement of early postural control in patients with PD.

Key words: Postural balance, Parkinson's Disease, outcome measures, rehabilitation, neurological disorders, psychometrics

Introduction

Impairment of postural control, leading to postural instability (i.e. inability to maintain or change posture in motor activities such as standing or walking) is one of the cardinal features of Parkinson's Disease (PD) and one of the main determinants of falls in this condition[1,2]. As a consequence, it is estimated that about 60% of persons affected by PD experience at least one fall and up to 50% of patients with PD are recurrent fallers. As falls in PD are a main source of disability and reduced quality of life[1,3], early identification of likely fallers among newly diagnosed patients with PD is an important rehabilitation goal in this population.

The Berg Balance Scale (BBS), one of the most widely used clinical scales for assessing balance[4], has been recommended as a valid and responsive tool for measuring impairment of postural control possibly leading to falls in PD[5,6]. However, it has also been suggested that the BBS, when used to measure balance in patients with PD, may: a) not cover completely, with its item content, the whole spectrum of postural control impairments typical of this condition[7,8]; b) display a ceiling effect[8]; and c) need revision of its rating scale structure at the item level[7]. In particular, Franchignoni and Velozo suggested that a Rasch analysis be performed on the BBS in order to assess and improve the scale's internal construct validity in reference to the PD population[7].

Rasch analysis is a powerful psychometric tool for assessing the internal construct validity of a scale, in view of the operationalization of the formal axiom of additive conjoint measurement performed by the mathematical model (i.e. the Rasch model), upon which it is based[9]. If an adequate fit of the scale to the Rasch model can be demonstrated within the context of Rasch analysis, the scale's total score can be transformed to an interval scale of measurement of ability. This is a tremendous advantage as it allows, unlike ordinal scales, the correct interpretation of change scores and proper access to parametric statistics, as

required in clinical trials[10].

The internal construct validity of the BBS has been assessed in a variety of settings and/or clinical conditions[4,11-13], but not so far in PD. Furthermore, to date no comprehensive assessment of the content validity of the BBS has been carried out. Thus, we set two main goals for the current study: i) to evaluate the content validity of the BBS, and ii) to evaluate, using Rasch analysis, the internal construct validity, reliability and targeting of the BBS in a sample of PD patients.

Materials and methods

Patients and setting

Data were collected prospectively at [redacted] from [redacted] to [redacted]. All subjects were consecutively enrolled before commencing, as outpatients, a rehabilitation program in a PD Clinic. The main inclusion criterion was a diagnosis of Parkinson's Disease, made by a neurologist according to the United Kingdom PD Society Brain Banking criteria[14]. Exclusion criteria were: cognitive impairment, as defined by a Mini-Mental State Examination (MMSE) score below 24/30; any other neurological or orthopedic disorder sufficiently severe to interfere with balance assessment. All patients were assessed in the morning, 60–120 min after their first morning drug intake, usually corresponding to a time of good performance even for patients with clinical fluctuations or dyskinesias.

All patients gave their informed consent to take part in the study, which was conducted in accordance with the ethical principles set forth in the Declaration of Helsinki[15].

Instruments

Enrolled subjects were assessed with the following tools:

1. The Unified Parkinson's Disease Rating Scale (UPDRS) part II 'Activities of Daily Living' (UPDRS-ADL) and part III 'Motor Examination' (UPDRS-ME)[16];

2. The Modified Hoehn and Yahr Scale (MHYS)[17], here used to characterize the sample in terms of disease progression. According to this tool, the first 3 disease stages (1, 1.5, and 2) are characterized by the absence of impairment of postural control, whereas increasing levels of the latter can be observed from stage 2.5 to 5.

3. The Berg Balance Scale (BBS)[18,19], a 14-item summative ordinal scale evaluating sitting balance, postural changes from sitting to standing and vice versa, transfers, and a variety of other standing balance tasks. Each item is scored from 0 to 4, where 0 implies the absence of balance ability and 4 the best possible performance in the observed activity. Thus, the total score ranges from 0 (lowest balance ability) to 56 (highest balance ability). The scale was administered by 4 licensed physiotherapists, all trained in administering the tool.

Assessment of the content validity of BBS

Content validity can be defined as the extent to which a measuring instrument contains items critical or appropriate to the construct being measured, i.e. the items cover substantively all the main theoretical aspects of the construct[20]. In order to assess the content validity of the BBS, we linked its items to both a general functioning and a balance-specific conceptual model. In order to appraise which domain(s) and categories of human functioning were addressed by the BBS, we linked its items to the 2nd level categories of the International

Classification of Functioning (ICF)[21] using standard linking techniques[22,23]. As the ICF does not provide specific categories for balance, we also linked the BBS items to the conceptual domains for postural control as suggested by Horak et al[24].

Rasch analysis

Following the above analysis, the BBS patient data were fitted to the Rasch Model[25]. This unidimensional mathematical model expects that the data fit a probabilistic Guttman pattern, so that a subject with a certain ability (level of balance) on the latent variable (i.e. balance) has a higher probability to affirm (i.e. pass) items requiring less ability in terms of balance (i.e. easier tasks), and a lower probability to affirm (i.e. fail) items requiring a higher ability level (i.e. more difficult tasks). The process of testing statistically whether the data fit the Rasch model's expectations and assumptions, here based on the partial credit parameterization of the model, is widely known as Rasch analysis, that has been reported in detail elsewhere[25-30]. Briefly, within the Rasch analysis, the following steps are sequentially performed:

1) To check whether the data fit to the Rasch model, by verifying: a) if there are significant deviations from the model's expectations among the item- and person-residuals (i.e. the standardized sum of all differences between observed and expected values summed over all persons for items and over all items for persons[27,29,30]), and at the level of the item characteristic curves (ICC), which show the difference between the observed and the expected responses predicted by the model on the basis of the probabilistic relationship between person ability and item difficulty for each item[25,30]; b) if the items maintain their stochastic ordering along the whole latent trait (item homogeneity or invariance)[25,29,30], and if the resulting item hierarchy is consistent with clinical expectations[30].

2) To check if the data satisfy the following model's requirements[31]:

   a) Monotonicity, i.e. the probability of endorsing an item response indicative of higher ability (e.g. better balance) should increase as the underlying level of the latent trait (balance) increases[29,31].

   b) Local independence, i.e. all the variation among responses to an item is accounted for by the person ability and, therefore, for the same value of ability there is no further systematic relationship among responses[31,32].

   c) Unidimensionality, i.e. all items measure a single underlying construct[29,31,33].

   d) Invariance at the subgroup level, also known as absence of differential item functioning (DIF) or item bias, that occurs when an item, regardless of displaying invariance at the whole sample level, shows a lack of invariance, i.e. DIF, across relevant subgroups (or person factors), such as gender or age. In this case, different groups of persons within a person factor respond in a different manner on the basis of their group membership, despite equal levels of the underlying characteristics[26,27,29,31].

3) To check the measurement quality, in terms of reliability of the scale[29,30,34,35], of test information[35,36], and of targeting of the scale to the sample[29,30,37,38].

4) To actively modify the data, on the basis of the evidences collected in the previous steps, should the data do not fit the Rasch model or do not meet the model's requirements. This is usually achieved by undertaking an iterative phase involving item modifications, aimed at finding a solution that satisfies both the model's expectations and assumptions, as well as the theoretical expectations about the

construct being measured[29]. These scale-modification strategies include:

a) item rescoring, to account for violations of the monotonicity requirement[29,30,39,40];

b) item grouping or 'testlets' creation, to account for violations of local independence[28,41,42];

c) Item splitting, to account for uniform-DIF[26,29];

d) Item deleting, a strategy to be adopted if all other strategies have failed and an item still displays misfit to the model or is biased by non-uniform DIF[42].

After each modification, model fit and satisfaction of the model's requirements are reassessed. This process is repeated cyclically, until no further modifications are needed and/or possible. Should a final solution fitting the model be found following the above modifications, in view of the statistical properties acquired by the total score (i.e. specific objectivity and sufficiency), the latter could be transformed into interval-level measurement, whose unit of measurement is the logit[25,29,30].

A full description of the methods used to assess model fit, model's requirements and measurement quality within the current Rasch analysis is available as an on-line supplementary table (S1).

[Table S1]

Statistical notes, softwares and sample size issues

All descriptive statistics were performed using SPSS software (SPSS. Version 13 for Windows; Release 13.0.1. SPSS Inc; 2004 (www.spss.com). Rasch analysis was carried out

on the whole data set using RUMM2030 software (version 5.4 for Windows. RUMM Laboratory Pty Ltd, Perth, Australia: 1997-2010 (www.rummlab.com).Within the context of Rasch analysis, a sample size of 285 observations would be sufficient to estimate item difficulty, with α of 0.01 to$<\pm$ 0.5 logits, irrespective of the targeting of persons to the items[43]. A significance value of 0.05 was used throughout and corrected for the number of tests by Bonferroni correction[44].

Results

Patients enrolled

All data were collected from a convenience sample of 285 consecutively enrolled individuals with Parkinson's Disease. Their demographic and clinical characteristics are summarized in table I.

[Table I]

Content validity of the BBS

As shown in table II, all BBS items could be linked to 2nd level categories pertaining to the 4th chapter (*Mobility*) of the *activities and participation* domain (d) of the ICF: 1 item (BBS05) was linked to *d420 (transferring oneself),* 3 items (BBS02, BBS03, BBS06) to *d410 (maintaining a body position),* whereas the remaining items were linked to *d415 (changing a body position)*.

Considering the conceptual model for postural control proposed by Horak et al., only 1 item (BBS03, "Sitting unsupported") could be linked to domain II (*limits of stability/verticality*); 4 items (BBS02, BBS06, BBS07, BBS12) were linked to domain V (*sensory orientation*), whereas all the remaining items were linked to domain III (*anticipatory postural*

adjustments/transitions). Thus, the analysis of content validity suggested that the BBS items assess these 3 main domains for postural control within the context of mobility activities requiring balance.

[Table II]

Rasch analysis

The base analysis performed on the 14 BBS items (Table III, analysis 1) showed that, although there were no individual misfitting items or persons, the scale failed the assumptions of monotonicity (50% of the items had disordered thresholds), local independence (there were 5 pairs of items with residual correlations above the LDRC, here set at 0.142), unidimensionality (PST=7.6%; LBCI: 5.1%) and invariance at the whole scale level (χ^2_{42}=57.705; p=0.054>0.03571). Furthermore, the scale appeared to be off-target, as the targeting index[38] was almost 5 standard errors of measurement (SEM) above the average item measure (set by default at 0 logits).

[Table III]

In the next stage of the analysis (Table III, analysis 2), we accounted for the violation of monotonicity by rescoring all items with disordered thresholds, according to clinically meaningful criteria made specific for each individual item (table IV). Monotonicity for item BBS03 could be achieved only through dichotomization (00001). After this rescoring (table III, analysis 2), the scale still failed the assumptions of local independence as there were three clusters of locally dependent items with residual correlations above the LDRC, set for this analysis at 0.132: BBS01-BBS04-BBS05 (postural changes and transfers; average residual correlation: 0.305), BBS13-BBS14 (standing with restricted base of support;

residual correlation: 0.221), and BBS02-BBS06-BBS07 (standing under various sensory orientation conditions; average residual correlation: 0.152). Furthermore, one item (BBS10) failed the requirement of stochastic invariance (χ^2_3=15.875; p=0.001<0.003571), although the item set satisfied the requirements of invariance at whole-sample level (χ^2_{42}=57.123; p=0.059>0.003571) and of unidimensionality (PST=5.8%; LBCI: 3.2%).

[Table IV]

As a consequence, we decided to account for local dependency by creating one testlet for each cluster of items whose residual correlation was above the LDRC. However, as the LDRC decreased by analysis, we created one testlet at a time, proceeding from the cluster of items with the highest dependency to that with the lowest dependency and reassessing, after each analysis, that the next candidate cluster was still locally dependent before merging it into a new testlet. Following this approach, we created 3 testlets, one for each of the above mentioned clusters of items, as detailed in Table III, analyses 3-5.

After these modifications, the scale appeared to satisfy all the model's requirements, at the item, the person, and the whole test level. However, the scale was still off-target (the average person measure was 4.303 SEM above the average item measure), and the targeting graph showed that the only threshold pertaining to BBS03 was 1.722 logits below the lowest person location (-4.677 logits). Furthermore, the visual inspection of the item characteristic curves (ICC) of the same item revealed a very flat curve for the observed responses in comparison to the responses expected by the model. Considering these findings, BBS03 was considered severely mistargeted and underdiscriminating and, hence, it was deleted.

The final 13-item scale, called BBS-PD (Table III, analysis 6), satisfied all the model's assumptions in terms of monotonicity (no items had disordered thresholds), local

independence (no pairs of items had a correlation of residuals above the LDRC of 0.082), strict unidimensionality (PST=4.0%; lower bound BCI: 1.4%) and invariance, both at the whole sample (χ^2_{24}=39.693; p=0.023>0.006250) and at the subgroup level (no significant DIF for gender or age for any item).

All persons, as well as all items (Table V), fitted the model individually. Visual inspection of the ICC curves confirmed that all items were adequately fitting the model. The item hierarchy (Table V) was consistent with clinical expectations, as the easiest items were those related to maintaining a body position (BBS03, BBS02, and BBS06) and to performing postural changes and transfers (BBS01, BBS04, and BBS05), whereas the most difficult items were those related to standing with a reduced base of support (BBS13 and BBS14).

[Table V]

The person reliability of BBS-PD, expressed both as PSI and Cronbach's α, was, respectively, 0.894 and 0.903, both values indicating precision of measurement at the individual level[29]. Given the PSI, persons could be separated into 4.6 strata (i.e. the statistically distinct levels of balance ability that BBS-PD was able to reliably distinguish[34]). Although the targeting of the scale was improved by the deletion of BBS03 and the ceiling effect was negligible (3.5%), the targeting index was 2.862, thus indicating that, on average, the ability of the sample was largely above the average difficulty of the BBS-PD. Indeed, the targeting graph (Figure 1) showed that persons were spread across 10 logits, with most of the subjects in the sample located in the right half of the measurement continuum. Particularly, as shown in table VI, 84.9% of the sample was located within [0.0, +5.0] logits, whereas just 43 subjects were located within [-5.0, 0.0] logits. Furthermore, the median value of the test information (i.e. a measure of how precisely the person ability is estimated[35]) was rather lower in the higher ability range ([+2.5, +5.0] logits) than in the middle ability ranges (as

expected), although up to 79 persons (27.7% of the entire sample) were located in this range. The original BBS scores for this group ranged from 51 to 56.

[Figure 1]

[Table VI]

The total raw score of the BBS-PD ranged 0-46. On the basis of the item calibrations, it was possible to construct a conversion table to transform the BBS-PD's raw scores into interval measures of ability, whose measurement unit is the logit (Table S2).

[Table S2]

Discussion

In this study, we fully evaluated the content and internal construct validity, reliability and targeting of the BBS in a sample of patients with PD, using Rasch analysis. This analysis demonstrated that, with some modifications, the BBS had adequate internal construct validity and reliability for measuring PD patients as individuals. However, the scale was not well targeted[38] to the sample as a whole as the items were, on average, less difficult than the mean ability of the persons with PD. Furthermore, our analysis of content validity demonstrated that the BBS does not contain items that assess postural responses and stability.

In this paper, as in other works on balance scales[28,40], the analysis of content validity was made by linking the BBS items to some conceptual model of balance. We adopted both a general functioning and a balance-specific conceptual model, i.e. Horak's model, mainly to assess construct coverage of the BBS. The ICF linking process showed that the BBS

measures the capacity in activities requiring balance, but not balance as a 'function'. On the other hand, as Horak's model describes the various 'functions' of postural control, it was more difficult to link the BBS items to this kind of model. In other words, as 'balance' is a multidimensional latent variable that could be linked to 'functions', there is a limited amount of information about the specific pattern of postural control impairments that can be inferred from clinical scales, such as the BBS, that assess a different ICF domain, i.e. 'activities'.

The basic Rasch analysis showed a violation of the monotonicity requirement for half of the BBS' items. This issue was also flagged in some previous reports on Rasch analysis of the scale, although the rescoring pattern here adopted differs from those previously reported[4,11]. This may suggest that the pattern of threshold disordering may be influenced by sample- and/or setting-specific factors.

Another important causative factor for the lack of internal construct validity displayed by the BBS was the presence of local dependency in the data, a common finding in health outcome scales[4]. Violation of the local independence requirement was not reported in previous Rasch analyses of the BBS[4,11-13]. In fact, it appears that in most of the previous works[11-13] the local independence requirement was not tested at all. This is not surprising, as local dependency has been rarely reported and/or addressed in health outcome studies based on Rasch analysis[32,42]. On the other hand, in the only paper on Rasch analysis of the BBS where this requirement was assessed[4], local dependency was not observed using the frequently recommended absolute cut-off of 0.3 for flagging any significant item residual correlations. However, another Rasch analysis on a 40-item set, including the same BBS data and aimed at building a balance item bank, revealed the presence of local dependency amongst the BBS items using the very same 0.3 cut-off [28]. We believe that this could be explained considering that local dependency went probably undetected with the smaller BBS item set (14 items)[4],

whereas it emerged within the larger UBS item set (40 items)[28]. This explanation is coherent with Marais' findings on simulated data[32], as she found that absolute cut-points of <0.3 or even <0.2 were unable to detect local dependency for scales with less than 20 items, as in the case of the BBS. In line with this view, for the interpretation of the item residual correlations, we chose not to use an absolute cut-off, but relative cutoffs (varying by analysis), as recently recommended[32]. This 'conservative" approach proved to be successful in correcting the distortions to the measurement process caused by the violation of local independence, thus avoiding the need to delete some of the dependent items. Indeed, by creating 'testlets' we were able to retain all the dependent items, thus preserving the original item content, adhering, at the same time, to the strict Rasch model's prescriptions for attaining scientific measurement[42].

It is acknowledged that violations of local independence may occur because of trait dependency (a form of multidimensionality, where some of the variation among responses is accounted for by a further latent variable) or response dependence (two or more items are linked in some way, so that the response to an item is influenced by the response to a previous item)[29,32]. Although it may be difficult to distinguish between the two forms, Andrich and Marais have shown that the impact of these two violations on reliability and targeting is opposite: in the case of trait dependency, reliability and person ability spread are reduced, whereas in the case of response dependency they are artificially inflated. Considering the reduction of reliability and person ability spread observed after creating the 'testlets', it is likely that the cause of local dependency in our analyses was response dependency. This is also supported by the fact that the 3 identified locally-dependent clusters contained items conceptually similar to one another (postural changes and transfers; basic standing under various sensory conditions; standing with a restricted base of support), as demonstrated by the content validity analysis. On the other hand, it is not entirely possible to

exclude that some trait dependency, although marginal, may have occurred as well, considering the mild improvement of unidimensionality observed across the 3 local dependency analyses.

Previous studies indicated that the BBS03 item (static sitting posture) was problematic, because of a ceiling effect[45], lower correlation with other items and lower factor loadings on factor analysis[4,45]. These findings could be explained also considering our content validity analysis, that showed that BBS03 differed substantially in content from the other items, as it was the only one that was linked to domain II (*limits of stability/verticality*) of the Horaks' model. Our Rasch analysis confirmed the marked ceiling effect of this item which was, as a consequence, severely underdiscriminating. Although the BBS03 item may be useful for very low ability patients with likely disturbance of trunk control, as in stroke[12], the lower categories of this item are unlikely to be endorsable by ambulatory patients, such as those in our sample. Although the deletion of an item should be considered a last resort for an already published scale[42], we deleted BBS03 because keeping this item would have added nothing to the precision of the person parameter estimates, given that off-target items add very little information in the population tested[35]. Furthermore, the deletion of BBS03 had no negative impact on the model fit and improved the overall targeting[38], especially at the left-end of the measurement range.

The study sample was made of patients that were consecutively enrolled, regardless of the disease stage, before commencing rehabilitation. Thus, it reflects the usual population of PD patients seen in our Rehabilitation Institution. According to the MHYS, in our sample there was a prevalence of persons in the 2.5th and 3rd stages, and this was entirely expected, given that in these stages postural control impairments are clinically manifest. However, according to the MHYS, 25% of the sample was not expected to show any signs of impaired postural

control (1st , 1.5th, and 2nd stages). Indeed, the fact that the BBS-PD ceiling effect was only 3.5% suggests that the MHYS was somehow inaccurate in detecting persons with initial signs of postural control impairment, thus confirming a finding already reported in the literature[46,47].

On the other hand, the fact that 27.7% of the sample was located in the upper ability quarter of the measurement range, where less precision of measurement is expected, suggests that in this critical range, person with different degrees of postural control impairments, may get similar BBS-PD scores. This finding may be explained substantively considering the evidence provided by the analysis of content validity, regarding the lack of items assessing those specific postural control impairments (i.e. postural reactions to slips and tripping and/or freezing during walking and/or turning[7,8]) that are typical of PD, and that may occur early in the course of the disease, even in highly functional individuals[8]. In summary, although the BBS (and its Rasch-modified counterpart) was able to detect clinical signs of postural control impairment better than the MHYS, its reduced precision in the higher functioning range may lead to underestimating the degree of postural control impairment for a given individual. This, in turn, may lead to an underestimation of the risk of falling in fully ambulatory patients, and to reduced responsiveness to interventions aimed at improving the early postural control impairments occurring in PD.

As this study was conducted on a convenience sample representing a cross-section of patients with PD drawn from a single rehabilitation center, the possibility of generalization of these findings to similar samples may be limited. Furthermore, although the sample size was large enough for stable calibrations, we did not have enough cases to draw a validation sample which would have enabled us to confirm the findings on the calibration sample. As a consequence, there is a risk concerning the obtained solution that its fit to the Rasch model

may be the result of chance. This should be borne in mind if using the raw-score interval scale transformation for the BBS-PD, that should be considered provisional at this time. Given these limitations, our findings require replication in the context of a larger multicenter study.

Conclusions

This study supports the internal validity and reliability of the BBS-PD as a measurement tool for patients with PD within the Rasch analysis framework. Although the BBS-PD may detect some degree of postural control impairment even in patients in the early stages of PD better than the Modified Hoehn & Yahr Scale, the lack of items assessing postural responses and stability during gait, that are typically impaired early in the course of the disease, hinders the precision of measurement of the tool in highly functional individuals with PD. This may lead to an underestimation of the risk of falling for a given individual, and to a reduced responsiveness to interventions aimed at improving early postural control impairments. These findings raise concerns regarding the effectiveness of the BBS as a measurement tool for the measurement of early postural control impairments in patients with Parkinson Disease.

REFERENCES

1. Kim SD, Allen NE, Canning CG, Fung VS. Postural instability in patients with Parkinson's disease. Epidemiology, pathophysiology and management. CNS Drugs 2013;27:97-112.
2. Latt MD, Lord SR, Morris JG, Fung VS. Clinical and physiological assessments for elucidating falls risk in Parkinson's disease. Mov Disord 2009;24:1280-1289.
3. Allen NE, Schwarzel AK, Canning CG. Recurrent falls in Parkinson's disease: a systematic review. Parkinsons Dis 2013;2013:906274.
4. La Porta F, Caselli S, Susassi S, Cavallini P, Tennant A, Franceschini M. Is the Berg Balance Scale an Internally Valid and Reliable Measure of Balance Across Different Etiologies in Neurorehabilitation? A Revisited Rasch Analysis Study. Arch Phys Med Rehabil 2012;93 (7):1209-1216.
5. Qutubuddin AA, Pegg PO, Cifu DX, Brown R, McNamee S, Carne W. Validating the Berg Balance Scale for patients with Parkinson's disease: a key to rehabilitation evaluation. Arch Phys Med Rehabil 2005;86:789-792.
6. 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.
7. Franchignoni F, Velozo CA. Use of the Berg Balance Scale in rehabilitation evaluation of patients with Parkinson's disease. Arch Phys Med Rehabil 2005;86:2225-2226; author reply 2226.
8. King LA, Priest KC, Salarian A, Pierce D, Horak FB. Comparing the Mini-BESTest with the Berg Balance Scale to Evaluate Balance Disorders in Parkinson's Disease. Parkinsons Dis 2012;2012:375-419.
9. Perline R, Wright BD, Wainer H. The rasch models as additive conjoint measurement. Applied Psychological Measurement 1979;3:237-255.
10. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaire. J Rehabil Med 2001;33.
11. Kornetti DL, Fritz SL, Chiu YP, Light KE, Velozo CA. Rating scale analysis of the Berg Balance Scale. Arch Phys Med Rehabil 2004;85:1128-1135.
12. Straube D, Moore J, Leech K, Hornby TG. Item analysis of the berg balance scale in individuals with subacute and chronic stroke. Top Stroke Rehabil 2013;20:241-249.
13. Wong CK, Chen CC, Welsh J. Preliminary assessment of balance with the Berg Balance Scale in adults who have a leg amputation and dwell in the community: Rasch rating scale analysis. Phys Ther 2013;93:1520-1529.
14. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992;55:181-184.
15. 59th World Medical Association General Assembly. Declaration of Helsinki: ethical principles for medical research involving human subjects. 2008. Accessed on 31/01/2009 from http://www.wma.net/en/30publications/10policies/b3/.
16. Fahn S, Elton R. The UPRDS Development Committee. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden C, Calne D, Goldstein M, eds. Recent developments in Parkinson's disease, vol 2. Florham Park, NJ: Macmillan, 1987: 153–163.
17. Jankovic J, McDermott M, Carter J, et al. Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. Neurology 1990;40:1529-1534.

18. Berg K, Wood-Dauphinée S, Williams J, Gayton D. Measuring balance in the elderly: preliminary development of an instrument. Physiotherapy Canada 1989;41.
19. Berg KO, Wood-Dauphinee SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;83 Suppl 2:S7-11.
20. Kucukdeveci AA, Tennant A, Grimby G, Franchignoni F. Strategies for assessment and outcome measurement in Physical and Rehabilitation Medicine: An educational review. J Rehabil Med 2011;43:661-672.
21. World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva2001.
22. Cieza A, Brockow T, Ewert T, et al. Linking health-status measurement to the international classification of functioning, disability and health. J Rehabil Med 2002;34:205-210.
23. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lesson learned. J Rehabil Med 2005;37:212-218.
24. Horak FB, Wrisley DM, Frank J. The Balance Evaluation Systems Test (BESTest) to differentiate balance deficits. Phys Ther 2009;89:484-498.
25. Andrich D. Rasch models for measurement. London: Sage Publications., 1988.
26. Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004;42:I37-48.
27. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1-18.
28. La Porta F, Franceschini M, Caselli S, Cavallini P, Susassi S, Tennant A. Unified Balance Scale: an activity-based, bed to community, and aetiology-independent measure of balance calibrated with rasch analysis. J Rehabil Med 2011;43:435-444.
29. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358-1362.
30. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods2009.
31. Kreiner S. The Rasch model ofr dichotomous items. In: Christensen KB, Kreiner S, Mesbah M, eds. in Rasch Models in Health. London UK, Hoboken NJ: ISTE Ltd and John Wiley & Sons, Inc, 2013.
32. Marais I. Local Dependence. In: Christensen KB, Kreiner S, Mesbah M, eds. in Rasch Models in Health. London UK, Hoboken NJ: ISTE Ltd and John Wiley & Sons, Inc, 2013.
33. Smith E. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement 2002;3:205–231.
34. Wright BD, Masters GN. Rating Scale Analysis. Chicago: MESA Press, 1982.
35. Kreiner S, Christensen KB. Person parameter estimation and measurement in Rasch Models. In: Christensen KB, Kreiner S, Mesbah M, eds. in Rasch Models in Health. London UK, Hoboken NJ: ISTE Ltd and John Wiley & Sons, Inc, 2013.
36. Salzberger T. Item Information: When Gaps Can Be Bridged. Rasch Measurement Transactions 2003;17:1:910-911.
37. Baghaei P. The Rasch Model as a Construct Validation Tool. Rasch Measurement Transactions 2008;12:1:1145-1146.
38. Fisher WPj. Rating Scale Instrument Quality Criteria. Rasch Measurement Transactions 2007;21:1:1095.

39. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3:85-106.
40. Franchignoni F, Horak F, Godi M, Nardone A, Giordano A. Using psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. J Rehabil Med 2010;42:323-331.
41. Wainer H, Kiely G. Item clusters and computer adaptive testing: A case for testlets. J Educ measurement 1987;24:185-202.
42. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. J Rehabil Med 2011;43:884-891.
43. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994;75:127-132.
44. Bland J, Altman D. Multiple significance tests: the Bonferroni method. British Medical Journal 1995;310:170.
45. Ottonello M, Ferriero G, Benevolo E, Sessarego P, Dughi D. Psychometric evaluation of the Italian version of the Berg Balance Scale in rehabilitation inpatients. Europa Medicophysica 2003;39:181-189.
46. Tanji H, Gruber-Baldini AL, Anderson KE, et al. A comparative study of physical performance measures in Parkinson's disease. Mov Disord 2008;23:1897-1905.
47. Goetz CG, Poewe W, Rascol O, et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations. Mov Disord 2004;19:1020-1028.

TITLES OF TABLES

TITLES OF FIGURES

Persons (n=285) and item thresholds are displayed, respectively, in the upper and the lower part of the graph, separated by the logit scale. Grouping set to interval length of 0.20 making 50 groups.

Tables

Table S1 (supplementary)– Assessment of model fit, model's requirements and measurement quality within Rasch analysis

Parameter	Test	Expected values / findings
Model fit: individual persons and items		
Person fit[a]	Fit Residual	Between -2.5 and +2.5
Item fit[a]	Fit Residual	Between -2.5 and +2.5
Item-trait interaction[b]	Chi-square	Non-significant (Bonferroni corrected)
Item Characteristics Curve (ICC) [c]	Visual inspection	Observed probabilities (class intervals) should match expected probabilities
Model fit: summary for persons and items		
Persons fit residual mean[d]	Mean of item fit residuals	0
Persons fit residual SD[d]	SD of item fit residuals	1
Items fit residual mean[d]	Mean of person fit residuals	0
Items fit residual SD[d]	SD of person fit residuals	1
Total item-trait interaction[e]	Chi-square	Non-significant (Bonferroni corrected)
Item hierarchy: face validity[f]	Visual inspection	Item hierarchy conforms to theoretical expectations
Model's requirements		
Monotonicity[g]	Visual inspection of each item's thresholds	All thresholds ordered
Local independence[h]	Correlation amongst items' residuals	<Local Dependency Relative Cut-off
Unidimensionality[i]	paired t-test on PCA of residuals	PST<5% or lower bound confidence interval (LBCI) PST<5%
Absence of Uniform-DIF[j]	Two way ANOVA	Main effect is non-significant (Bonferroni corrected)
Absence of Non-Uniform-DIF[j]	Two way ANOVA	Interaction effect is non-significant (Bonferroni corrected)
Measurement quality: reliability and targeting		
Person reliability: separation[k]	Person Separation Index (PSI)	≥0.70 for group measurement; ≥0.85 for individual person measurement
Person reliability: strata[l]	Number of strata	≥2 strata for group measurement; ≥3 strata for person measurement
Person reliability (CCT)[m]	Cronbach's alpha	≥0.70 for group measurement; ≥0.85 for individual person measurement
Person reliability: measurement error[n]	Standard Error of Measurement (SEM)	Expected to be as low as possible
Precision of the person measure estimates[o]	Test information	Expected to be higher in the middle of the measurement range
Targeting[p]	Targeting Index	[-1, 1]: good targeting; [-2, 2]: fair targeting
Ceiling effect[q]	Calculation of % of persons with maximum score	<2%
Floor effect[q]	Calculation of % of persons with minimum score	<2%

NOTES.

[a] Fit to the model for individual items and persons is expressed by fit residuals (FR), respectively for items and persons. FR are the standardized sum of all differences between observed and expected values summed over all persons for items and over all items for persons[29,30]. Fit residuals are expected to be 0 in case of achievement of a perfect stochastic Guttman pattern. Negative values suggest that the observed responses tend to be deterministic (i.e. lacking of the expected randomness), whereas positive values suggest that the observed response tend to follow less the expected Guttman pattern (i.e. excessively random). In case of adequate fit to the model, they are expected to be in the range [-2.5, 2.5], which represents the 99% confidence interval around the fit residual[27,30]. Values outside this confidence intervals suggest misfit.

[b] Item-trait interaction refers to the requirement of homogeneity or item invariance[31], i.e. the items should maintain their stochastic ordering along the whole latent trait[25,29]. The χ^2 statistics tests the homogeneity (invariance) assumption by comparing the difference between in expected values and observed values across groups representing different ability levels (called class intervals) and across the trait to be measured. If significant, taking into account a Bonferroni-corrected p-value, a violation of the invariance (homogeneity) requirement of the item hierarchy is suggested[30].

[c] The Item Characteristic Curve (ICC) displays the expected probabilities to pass the item for any ability level along the measurement continuum. In order to assess fit to the model, another curve is constructed by connecting the observed probabilities values across the trait (represented by the various class intervals). The match between the two curves is then assessed[30]. A good match between the two curves suggest normal discrimination. A flatter observed probability curve suggests that the item is under-discriminating, i.e. the responses to the item are too erratic and do not follow the Rasch model's expectations[30]. On the other hand, a steeper curve suggests over-discrimination, i.e. the responses to the item lacks the expected randomness and tend to be too deterministic[30]. Independently from item discrimination, individual class interval markedly outside the expected probability curve suggest that responses to this item, at some ability levels, deviate from the model's expectations, thus suggesting violation of the homogeneity (invariance) requirement[25, 30].

[d] Fit to the model for individual item and persons can be summarized as a mean and a standard deviation of the fit residuals, respectively for items and persons. In case of perfect fit to the model, such means and standard deviations are expected to assume values equal to 0 and, respectively, 1, as they are transformed to approximate a z score, representing a standardized normal distribution[27].

[e] This total chi-square is calculated by summing up the chi-squares of the individual items (please see note [b]) divided by the sum of their degrees of freedom minus 1[29,30]. As for the item-trait interaction chi-square for individual items, a violation of the invariance (homogeneity) requirement of the item hierarchy is suggested if this summary chi-square is significant (taking into account a Bonferroni-corrected p-value).

[f] The difficulty order of the items, suggested by the analysis, should make sense from a clinical point of view and be consistent with the expectations derived from theory. If this is the case, it provides evidence towards the construct validity of the item set with regard to the variable being measured[30].

[g] The probability of endorsing an item response indicating higher ability (e.g. better balance) should increase as the underlying level of the latent trait (balance) increases (monotonicity requirement)[31]. As a consequence, the difficulty thresholds (i.e. transition point between adjacent scoring categories) appear ordered. If the response options for a given item are used inconsistently (e.g. because of misinterpretation of the scoring options, caused by too many scoring options or by inaccurate labeling of the options), the difficulty thresholds appear disordered[29].

[h] All the variation among responses to an item is accounted for by the person ability and, therefore, for the same value of ability there is no further systematic relationship among responses (local independence requirement)[31]. Items are considered to be locally dependent if their residual correlation is above a Local Dependency Relative Cutoff (LDRC), calculated by adding 0.2 to the average of residual correlations, after having removed the correlation of each item to itself, equal to 1[32].

[i] All items measure a single underlying construct (unidimensionality requirement)[29,31]. Unidimensionality is tested post-hoc with a paired t-test on separate estimates for each respondent (derived from subsets of items identified by a principal component analysis of the item residuals)[33]. Unidimensionality is considered achieved when the PST (percentage of significant t-test) is <5% (strong unidimensionality), or the LBCI (lower bound of the binomial confidence interval for proportions) is <5% (acceptable unidimensionality)[29].

[j] Item bias or DIF occurs when an item, regardless of displaying invariance at the whole sample level, shows a lack of invariance, i.e. DIF, across relevant subgroups (or person factors), such as gender or age[29, 31]. In this case, different groups of persons within a person factor respond in a different manner on the basis of their group membership, despite equal levels of the underlying characteristics. The presence of DIF is tested by a two-way ANOVA for each item, where scores are compared across each level of the person factor and across different ability levels, as summarized by the class intervals (please see note [b])[27]. In case of Uniform-DIF (U-DIF), the item bias is systematic along the trait, as suggested by a significant main effect for the person factor[27]. In case of Non-Uniform-DIF (NU-DIF), the item bias varies along the trait, as suggested by a significant interaction effect (person factor × class interval)[27]. Significance p-values are Bonferroni-corrected.

[k] PSI is calculated as the ratio between the variance among the estimates of persons tested and the error variance for each person; it indicates how reliably the persons are separated[30]. 0.70 is considered the absolute minimum for group measurement, whereas 0.85 is considered the absolute minimum for individual person measurement[29].

[l] Strata are the number of statistically distinct levels of person ability (person strata) that the scale is able to reliably distinguish[34].

[m] Chronbach's alpha (Classical Test Theory reliability) is derived as the proportion of variance of the true score and the total variance including error[30].

[n] Closely related to the concept of reliability within the Classical Test Theory framework, is that of the Standard Error of Measurement (SEM), that is calculated as follows: $SEM = SD \times \sqrt{1-r}$, where SD is the standard deviation of the person measures, and r is the reliability coefficient (i.e. the Person Reliability Index within the Rasch analysis framework)[30]. It provides an indication of the dispersion of the measurement errors when trying to estimate person abilities from their observed scores. It is less meaningful within the Rasch analysis framework, as in the latter the standard errors are individually calculated for each person measurement[30]. However, it is here reported as it is used to calculate the targeting index (please see note [p])

[o] Test information (I)[35,37] for a given person estimate is calculated as follows: $I = \frac{1}{SE^2}$, i.e. the reciprocal of the squared standard error around the person measure. It is a measure of how precisely the person ability is estimated[36].

[p] The targeting index is calculated as the ratio between the Person location mean and the SEM (Standard Error of Measurement; please see note [n] for the formula)[39]. It gives an indication of how the average person location has moved away from the average item difficulty, set by default at 0 logits.

[n] Ceiling and floor effects provide an indication on how many persons in the sample have received the higher and, respectively, the lower score of the scale[39].

Table I – Demographic and clinical characteristics of the study patients (N=285)

	N	%	Median	Min	Max	Mean	StDev
Age (years)	285	-	72	41	86	71.2	7.0
Gender							
Males	130	45.6	-	-	-	-	-
Females	155	54.4	-	-	-	-	-
Motor fluctuations							
Present	151	52.9	-	-	-	-	-
Absent	134	47.1	-	-	-	-	-
UPDRS-ADL (total score)	285	-	16	0	32	16.7	5.9
UPDRS-ME (total score)	285	-	24	3	40	24.0	6.6

			BBS total scores				
	N	%	Median	Min	Max	Mean	StDev
Modified Hoehn and Yahr scale							
All stages	285	-	49	4	56	46.0	9.7
Stage 1	7	2.5	54	53	55	54.0	1.0
Stage 1.5	37	13.0	55	39	56	52.8	4.4
Stage 2	27	9.5	52	39	56	50.8	4.6
Stage 2.5	69	24.2	50	32	55	48.8	5.6
Stage 3	109	38.2	43	14	56	42.6	9.7
Stage 4	32	11.2	36.5	21	46	34.2	8.3
Stage 5	4	1.4	4*	4	4	-	-

NOTE: *at stage 5 of the Modified Hoehn and Yahr, for only one patient it was possible to calculate the total score, as the other 3 ones had several item missing item data.

UPDRS-ADL: UPDRS version 3.0, part II (Activities of Daily Living); UPDRS-ME: UPDRS version 3.0, part III (Motor Examination).

Table II – Content validity of the BBS according to the ICF and Horak's model for postural control

BBS items	ICF model (2nd level category)	Horak's domains for postural control I	II	III	IV	V	VI
BBS01 - From sitting to standing	d410 Changing basic body position			*			
BBS02 - Standing unsupported	d415 Maintaining a body position					*	
BBS03 - Sitting unsupported	d415 Maintaining a body position		*				
BBS04 - From standing to sitting	d410 Changing basic body position			*			
BBS05 - Transfers	d420 Transferring oneself			*			
BBS06 - Standing with eyes closed	d415 Maintaining a body position					*	
BBS07 - Standing with feet together	d415 Maintaining a body position					*	
BBS08 - Reaching forward while standing	d410 Changing basic body position			*			
BBS09 - Retrieving object from floor	d410 Changing basic body position			*			
BBS10 - Turning trunk (feet fixed)	d410 Changing basic body position			*			
BBS11 - Turning 360°	d410 Changing basic body position			*			
BBS12 - Placing alternate foot on stool	d410 Changing basic body position					*	
BBS13 - Tandem standing	d410 Changing basic body position			*			
BBS14 - Standing on one leg	d410 Changing basic body position			*			

Abbreviations. BBS: Berg Balance Scale; ICF: International Classification of Functioning; I: Biomechanical constraints; II: Limits of stability / verticality; III: anticipatory postural adjustments / transitions; IV: postural responses; V: sensory orientation; VI: stability in gait

Table III – Summary of Rasch analysis for BBS-PD

	Item location		Person location				Item fit residuals		Person fit residuals		Item-trait interaction		Reliability			Unidimensionality t-test (CI)		
Analysis stage	*Mean*	*SD*	*Mean*	*SD*	*SEM*[a]	*Targeting index*[b]	*Mean*	*SD*	*Mean*	*SD*	*Chi-sq (df)*	*P*	*PSI with extremes*	*PSI without extremes*	*α with extremes*	*Number of significant t-tests*	*PST (%)*	*Lower bound 95% CI*
1 Base	0.0	2.838	2.787	1.930	0.579	4.813	-0.639	1.119	-0.041	0.151	57.705 (42)	0.054	0.910	0.915	0.934	21 out of 275	7.6	5.1
2 After rescoring	0.0	2.425	2.697	2.005	0.588	4.587	-0.657	1.092	-0.039	0.142	57.123 (42)	0.059	0.914	0.920	0.930	16 out of 275	5.8	3.2
3 After creating testlet 1	0.0	2.559	2.550	1.914	0.574	4.441	-0.501	1.096	-0.036	0.123	61.842 (36)	0.005	0.910	0.914	0.907	14 out of 275	5.1	2.5
4 After creating testlet 2	0.0	2.547	2.715	1.834	0.559	4.854	-0.531	1.158	-0.035	0.122	59.319 (33)	0.003	0.907	0.913	0.888	13 out of 275	4.7	2.2
5 After creating testlet 3	0.0	2.595	2.363	1.703	0.549	4.303	-0-501	1.025	-0.043	0.139	40.376 (27)	0.047	0.896	0.902	0.889	11 out of 275	4.0	1.4
6 After deleting BBS3	0.0	1.041	1.559	1.673	0.545	2.862	-0.399	1.140	-0.227	0.823	39.693 (24)	0.023	0.894	0.895	0.903	11 out of 274	4.0	1.4
Recommended values	*0.0*		*0.0*			*[-1, 1] [-2, 2]*[b]	*0.0*	*1.0*	*0.0*	*1.0*		*>0.006*[c]		*>0.850 >0.700*[d]			*<5.0*[e]	*<5.0*[e]

NOTE. Values are mean ± SD or as otherwise indicated.

Abbreviations: BBS-PD: Berg Balance Scale for Parkinson's disease; SD, standard deviations; df, degrees of freedom; P, bonferroni-corrected χ^2 probability value; PSI, person separation index; PST, percentage of significant t-test carried out on the estimates that, within a principal component analysis of residuals, loaded positively and negatively (factor loading >±.3) on the first component; CI, binomial confidence interval for PST.

Testlet 1, BBS01-BBS04-BBS05 (postural changes and transfers). Testlet 2, BBS13-BBS14 (standing with restricted base of support). Testlet 3, BBS02-BBS06-BBS07 (standing).

[a] SEM is the Standard Error of Measurement of the person locations, calculated with the formula: $SD \times \sqrt{1 - reliability}$, where SD is person location standard deviation and reliability is the PSI with extremes.

[b] The targeting index is calculated as the ratio between the average person measures and the SEM. Targeting is good, and respectively fair, when the average person measure is beyond [-1 +1] and, respectively, [-2, +2] SEM the average item measure (set by default at 0 logits).

[c] Bonferroni-corrected value of .05, indicative of statistical significance, will vary by analysis; the reported value refers to the final solution.

[d] A value of >.850 indicates precision of measurement also at the individual level, whereas a value of >.700 indicates precision only at the group level.

[e] Unidimensionality is considered achieved either when PST is <5% or when the lower bound of its binomial CI is <5%.

Table IV – New rating scale scorings for the BBS-PD items (analysis no. 2)

Item	Rescored	Max Score	Scoring model 1	2	3	4	5
BBS 01-From sitting to standing	-	4	0	1	2	3	4
BBS 02-Standing unsupported	Y	3	0	1	1	2	3
BBS 03-Sitting unsupported	Y	1	0	0	0	0	1
BBS 04-From standing to sitting	Y	3	0	1	1	2	3
BBS 05-Transfers	-	4	0	1	2	3	4
BBS 06-Standing with eyes closed	Y	3	0	1	1	2	3
BBS 07-Standing with feet together	Y	3	0	1	1	2	3
BBS 08-Reaching forward	-	4	0	1	2	3	4
BBS 09-Retrieving object from floor	Y	3	0	1	1	2	3
BBS 10-Turning trunk (feet fixed)	-	4	0	1	2	3	4
BBS 11-Turning 360°	-	4	0	1	2	3	4
BBS 12-Placing alternate foot on stool	-	4	0	1	2	3	4
BBS 13-Tandem standing	Y	3	0	1	1	2	3
BBS 14-Standing on one leg	-	4	0	1	2	3	4

Abbreviation: BBS-PD: Berg Balance Scale for Parkinson's disease

NOTES. For each item the rescoring pattern is presented. For instance, for item BBS05 no changes were made to its scoring model (01234), whereas for item BBS03, the first four categories were collapsed together, while the last category remained unchanged (00001)

Table V − Item parameters, fit statistics and content validity for the BBS-PD items (analysis no. 6)

	Model fit					Content validity: ICF	Content validity: Horak's model					
BBS items	**Location**	**SE**	**FitResid**	**χ^2**	**Prob***	**(2^{nd} level category)**	**I**	**II**	**III**	**IV**	**V**	**VI**
02-06-07	-1.355	0.073	-1.027	4.696	0.195	d415 Maintaining a body position					*	
01-04-05	-1.118	0.058	0.390	7.715	0.052	d410 Changing basic body position; d420 Transferring oneself			*			
09	-0.713	0.110	-1.799	8.010	0.046	d410 Changing basic body position			*			
10	-0.395	0.089	1.369	6.136	0.105	d410 Changing basic body position			*			
08	0.341	0.093	0.525	1.934	0.586	d410 Changing basic body position			*			
12	0.739	0.071	-1.844	7.485	0.058	d410 Changing basic body position					*	
11	1.094	0.078	-0.564	2.917	0.405	d410 Changing basic body position			*			
13-14	1.408	0.053	-0.247	0.798	0.850	d410 Changing basic body position			*			

NOTES. BBS-PD items (or testlets) are ordered by progressively increasing difficulty from top to bottom. The location is expressed in logits. The degrees of freedom for each χ^2 were 3 for all items. Item BBS03 was subsequently deleted. Item BBS10 has a significant χ^2, indicative of some misfit to the model. Finally, the linked categories of the ICF and Horak's model for postural control are displayed for each item. *The Bonferroni-corrected p-value indicating statistical significance at the .05 level was .003.

Abbreviations. BBS-PD: Berg Balance Scale for Parkinson's disease; SE, standard error; FitResid, fit residual; Prob, χ^2 probability. I: Biomechanical constraints; II: Limits of stability / verticality; III: anticipatory postural adjustments / transitions; IV: postural responses; V: sensory orientation; VI: stability in gait.

Table VI – Test information by BBS-PD measurement strata

Stratum	Person location range (logits)	N	% of the sample	Test information (median value)
1	[-5.0, -2.5]	6	2.1	3.663
2	[-2.5, 0.0]	37	13.0	6.643
3	[0.0, +2.5]	163	57.2	6.747
4	[+2.5, +5.0]	79	27.7	2.058

NOTES. The measurement continuum ([-5.0, +5.0] logits) was divided in 4 different ranges, each spanning 2.5 logits, on the basis of the number of estimated reliability strata. For each of those strata, the number of patients and the median test information for that range are presented. Information is a measure of how precisely the person ability is estimated by the item set. For a well targeted scale, it is expected that most persons in the sample will be located in the middle strata (where precision of measurement is higher), with fewer subjects in the peripheral strata, where precision of measurement (and, thus, information) is reduced. As the BBS-PD is not well targeted, the reduced precision of the person estimates affects a significant percentage of the sample in the higher ability (i.e., the 4th) stratum.

Table S2 (Supplementary) – Raw score to measure transformation table [can be published as supplementary online material]

Raw score	Measure (logit)	
	Location	Std Error
0	-4.934	1.168
1	-4.104	0.797
2	-3.600	0.605
3	-3.293	0.504
4	-3.084	0.449
5	-2.922	0.417
6	-2.782	0.399
7	-2.652	0.390
8	-2.521	0.389
9	-2.381	0.394
10	-2.226	0.404
11	-2.052	0.414
12	-1.866	0.421
13	-1.677	0.424
14	-1.491	0.422
15	-1.309	0.417
16	-1.133	0.411
17	-0.964	0.404
18	-0.801	0.396
19	-0.645	0.388
20	-0.496	0.379
21	-0.355	0.371
22	-0.221	0.363
23	-0.093	0.356
24	0.029	0.351
25	0.147	0.346
26	0.262	0.343
27	0.374	0.342
28	0.486	0.342
29	0.599	0.344
30	0.714	0.347
31	0.832	0.352
32	0.955	0.359
33	1.084	0.366
34	1.221	0.375
35	1.365	0.385
36	1.517	0.396
37	1.679	0.409
38	1.850	0.423
39	2.032	0.440
40	2.228	0.461
41	2.443	0.490
42	2.690	0.531
43	2.991	0.593
44	3.393	0.697
45	4.013	0.901
46	4.968	1.301

This conversion table is to be used only if patients are assessed on all the 13 BBS-PD items and if the modified rating scales for the items are to be used, as detailed in table IV.