RESEARCH PAPER

Reliability of four scales on balance disorders in persons with multiple sclerosis

DAVIDE CATTANEOR, JOHANNA JONSDOTTIR & STEFANIA REPETTI

LaRiCE: Gait and Balance Disorders Laboratory, Department of Neurorehabilitation, Don Gnocchi Foundation I.R.C.C.S., Milan, Italy

Accepted December 2006

Abstract
Purpose. Balance rehabilitation is an important component of the retraining program in people with multiple sclerosis (MS). Measuring balance is fundamental for an accurate assessment and therapy selection. The aim of this study was to compare interrater and test-retest reliability of the Berg Balance Scale, the Dynamic Gait Index, the Dizziness Handicap Inventory and the Activities-specific Balance Confidence.
Method. A group of 25 persons were enrolled in the study. The group consisted of 8 males and 17 females, mean age 41.7 years (12.5 years, SD). The onset of pathology was 8.7 years (8.8 years SD) before the beginning of the study. To assess the test-retest reliability two consecutive assessments were collected by the same rater. To assess the inter-rater reliability, persons were concurrently assessed by two raters.
Results. The Intraclass Correlation Coefficients (ICCs) for interrater reliability ranged between 0.94 and 0.96. The ICCs for test-retest reliability ranged between 0.85 and 0.96.
Conclusion. The inter-rater reliability of the instruments proved to be satisfactory. Lower but acceptable results were obtained for the test-retest paradigm. The data obtained in this study suggest that these scales are reliable tools for assessing balance function in persons suffering from MS.

Keywords: Reliability, balance, rehabilitation, multiple sclerosis

Introduction
Balance rehabilitation is an important component of the retraining programme in people with multiple sclerosis (MS) [1–3]. Measuring balance is fundamental to accurate assessments, appropriate therapy selection and the measurement of outcomes.
A variety of laboratory techniques and clinical scales have been proposed [4], however, clinical scales are the instruments most commonly used in clinical settings. They provide insight for the planning of rehabilitation programme, are less expensive than laboratory assessments and easily applied in the clinical setting.
In addition to scales measuring balance performance, the use of scales measuring the person’s perception of balance and the behavioural consequences of balance disorders are important to assess the impact of balance disorders on activities of daily life. The Berg Balance Scale (BBS) [5], the Dynamic Gait Index (DGI) [6], the Dizziness Handicap Inventory (DHI) [7] and the Activities-specific Balance Confidence (ABC) [8] have gained popularity within the clinical and scientific community. The BBS and the DGI are clinical scales that evaluate balance in standing and during gait activities respectively. They are well known among physiotherapists and allow comparisons among groups of persons with various pathologies. The ABC and DHI are questionnaires that measure perception of balance in activities of daily life (ADL) and perception of the level of handicap. Information gained from the two scales measuring balance perception may better account for reduction of daily activities than counting number of falls or measuring clinical balance performance. Taking into account...
behavioural aspects is important for addressing the person’s beliefs about his ability to carry out activities of daily life and tuning the person’s behaviour to his actual level of impairment and disability. Information from all four instruments would make it possible to develop customized rehabilitation programmes taking into account both physical impairments and the persons’ own perception of balance, as well as, behavioural consequences of both.

To be effective for clinical and/or research purposes the scales have to show good performance in three psychometric characteristics: Validity, reliability and responsiveness. Validity of these scales for a population with MS has been reported [9] but no study has reported their reliability for a population with MS.

The aim of this study was to compare reliability of the four balance scales in a cohort of persons suffering from MS.

Method

Subjects

Forty-five persons suffering from MS attending the Department of Multiple Sclerosis of Don Gnocchi Foundation were assessed.

To be eligible for the study, the persons had to meet the following inclusion criteria: Clinically or laboratory definite relapsing-remitting or secondary progressive multiple sclerosis, ability to stand independently in upright position more than 30 sec, ability to walk for 6 m with or without an assistive device. Cognitive impairment of all persons had been clinically assessed a priori by the medical staff. Persons with cognitive impairments that might hinder understanding of the tasks to be accomplished were excluded from enrolment. A sub-group of 25 persons met the inclusion-exclusion criteria and were enrolled in the study. The group consisted of 8 males and 17 females, mean age 41.7 years (12.5 years SD). The onset of pathology was 8.7 years (8.8 years SD) before the beginning of the study. Four persons used a walking aid in their daily activities. After informed consent was obtained, persons completed a questionnaire providing information about their age, the onset of pathology and other demographic characteristics. Then the battery of tests was administered.

To assess the test-retest reliability, two consecutive assessments were collected by the same rater. To avoid influences due to change in the pharmacological and physiotherapy treatments and behavioural consequences of hospital stay, the data were collected three days apart at the beginning of the hospital stay; no rehabilitative sessions, drugs or invasive exams were provided between the first and the second assessment. Self-administered tests were collected at both assessments. To reduce the person’s recall of the first assessment of the scores assigned to self-administered tests, the order of the items of the second test was varied. The tests were divided in to two parts and the second half was administered first and then the first half.

To assess the inter-rater reliability of the BBS and DGI scales persons were concurrently assessed by two raters who were physiotherapists with long experience in neurorehabilitation. Inter-rater reliability for the ABC and the DHI was not assessed because the scales are self-administered.

The assessment was carried out in one session, the testing protocol took 30–40 min to be administered, persons were allowed to rest during testing if necessary. Persons were tested wearing their normal shoes.

Because five of the auto-administered tests were not completed the sample for these scales consisted in 20 subjects.

Instruments

Berg Balance Scale (BBS). This scale rates performance from 0 (cannot perform) to 4 (normal performance) on 14 items. The items explore the ability to sit, stand, lean, turn and maintain the upright position on one leg [5]. The psychometric properties of the scale have been assessed on a population of elderly subjects. For that population of subjects, the scale proved to be a valid and reliable instrument. In another study on elderly subjects the test retest and interrater reliability of BBS was found to be very high with the ICCs ranging from 0.98–0.99 for intrarater reliability and 0.98 for interrater reliability [10]. In our previous study on the validity of the BBS for subjects with MS the scale was found to have good concurrent validity but not a good discriminant validity, that is, because of a low level of sensivity it did not distinguish well between fallers and non-fallers [9]. However, a cut-off score of 44 was established as a criterion to identify MS subjects with a high risk of fall based on results in the literature [9]. Performance on both the BBS and the DGI was rated by the first two authors (DC and JJ).

Dynamic Gait Index (DGI). This scale measures mobility function and dynamic balance. The eight tasks of this scale include walking, walking with head turns, pivoting, walking over objects, walking around objects and going up stairs [6,11]. The performance is rated on a 4-point scale. McConvey and Bennett [12] found an inverse correlation of 0.80 between the scores of the DGI and the time to walk 6.1 m in persons with MS. Data on reliability were published in the same study using videotaped recordings of the person’s performances. The study showed an
interrater reliability of 0.98 and test-retest reliability (Pearson’s correlation coefficients) ranging from 0.76–0.99. However, their study did not permit an evaluation of day-to-day variability of performance which is one of the reasons the present study was undertaken.

*Activities-specific Balance Confidence (ABC).* This is a scale in which the person rates his self-perceived level of confidence while performing 16 daily living activities [8]. In a study by Miller et al. [13] the test-retest reliability of the ABC Scale among people who have a lower-limb amputation was 0.91. Moreover, correlations with the 2 Meters Walking test and Timed Up and Go test were observed to be 0.72 and −0.70 [13]. Another study that aimed to assess convergent validity between the ABC Scale and Falls Efficacy Scale (FES) in a population of elderly people indicated that the ABC and FES were highly correlated [14].

*Dizziness Handicap Inventory (DHI).* This scale is a multidimensional self-assessment scale that quantifies the level of disability and handicap in three subscales: physical, emotional and functional [7]. Scores range from 0–100; where 100 means high level of disability and handicap. DHI demonstrated good internal consistency (0.91) and test retest reliability (0.97) [15]. Since in the other tests evaluated here higher scores mean better performances, the scores of the DHI were transformed: \( \text{DHI}_{\text{MODIFIED}} = 100 - \text{DHI} \), so 100 points indicates a high level of performance.

**Data analysis**

Total scores of the tests were used to compute absolute agreement Intraclass Correlation Coefficients (ICCs), using a two-way, random effects model. In that way raters are conceived as being a random selection from among all possible raters, and subjects are conceived as being a random factor too [16]. Standard errors of measurements for total scores were calculated for test-retest and interrater reliability [17]. Standard error of measurement of a test (for example in a test-retest condition), refers to the variability of test scores that would have been obtained from a single person had that person been tested multiple times. Change scores higher than \( 2 \times \text{SEM} \) points (for example 3 points in the BBS) cannot be due only to exposition to two consecutive assessments. Bland-Altman scatter plots [18] were used to illustrate the consistency of measurements and the presence of trends.

Since it is possible that the exposition to two consecutive executions of the test might improve subjects’ performance, we compared the first baseline assessment with the retest assessment to verify the presence of a learning effect.

**Results**

Three persons reported 1 fall and 1 person reported 10 falls in the month before the beginning of the study; the mean number of falls was 0.68 (2.29 SD). Table I shows the test scores for the whole group.

The group of subjects reported mild impairment in static balance (on the average reaching 84% of the maximum score of BBS scale) and more pronounced impairment in dynamic balance (on the average reaching 66% of the maximum score of DGI scale). The self-administered tests showed moderate (about 50% of the maximum score of the scales) confidence of persons in their balance skills.

The effect of two consecutive assessments accounted for less than 1 point of improvement on the retest both on the BBS and the DGI scales, those improvements corresponded to a 1.6% and 3.5% of improvement respectively for BBS and DGI. Self-administered scales did not show a learning effect. The scores decreased 4 points for the ABC scale and 1 points for the DHI scale, corresponding to a 4% and 1% of decrement respectively for ABC and DHI.

Table II shows the Intraclass Correlation Coefficients for the test-retest condition. The highest coefficient was observed for the scale rating static balance (BBS), conversely for the DGI the coefficient was lower. No relevant differences of ICC values were observed between self-administered scales.

Table III shows the Intraclass Correlation Coefficients for the interrater condition. In contrast to the test-retest condition where ICCs of BBS and DGI were different in the interrater condition the tests

<table>
<thead>
<tr>
<th></th>
<th>Descriptive statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n )</td>
</tr>
<tr>
<td>BBS – bas</td>
<td>25</td>
</tr>
<tr>
<td>BBS – Int</td>
<td>25</td>
</tr>
<tr>
<td>BBS – ret</td>
<td>25</td>
</tr>
<tr>
<td>DGI – bas</td>
<td>25</td>
</tr>
<tr>
<td>DGI – int</td>
<td>25</td>
</tr>
<tr>
<td>DGI – ret</td>
<td>25</td>
</tr>
<tr>
<td>ABC – bas</td>
<td>20</td>
</tr>
<tr>
<td>ABC – ret</td>
<td>20</td>
</tr>
<tr>
<td>DHI – bas</td>
<td>20</td>
</tr>
<tr>
<td>DHI – ret</td>
<td>20</td>
</tr>
</tbody>
</table>

BBS, Berg Balance Scale; DGI, Dynamic Gait Index; ABC, Activity-specific Balance Confidence; DHI, Dizziness Handicap Inventory. Bas, baseline assessment; first rater; Int, assessment of second rater; Ret, second assessment of the first rater. SEM, standard error of measurements.
Table II. Intraclass Correlation Coefficients for the test-retest condition.

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>CI+</th>
<th>CI−</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS</td>
<td>0.96</td>
<td>0.91</td>
<td>0.98</td>
</tr>
<tr>
<td>DGI</td>
<td>0.85</td>
<td>0.71</td>
<td>0.93</td>
</tr>
<tr>
<td>ABC</td>
<td>0.92</td>
<td>0.80</td>
<td>0.97</td>
</tr>
<tr>
<td>DHI</td>
<td>0.90</td>
<td>0.77</td>
<td>0.96</td>
</tr>
</tbody>
</table>

BBS, Berg Balance Scale; DGI, Dynamic Gait Index; ABC, Activity-specific Balance Confidence; DHI, Dizziness Handicap Inventory. ICC, Intraclass Correlation Coefficient; CI, 95% confidence intervals.

Table III. Intraclass Correlation Coefficients for the interrater condition.

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>CI+</th>
<th>CI−</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS</td>
<td>0.96</td>
<td>0.90</td>
<td>0.97</td>
</tr>
<tr>
<td>DGI</td>
<td>0.94</td>
<td>0.88</td>
<td>0.97</td>
</tr>
</tbody>
</table>

BBS, Berg Balance Scale; DGI, Dynamic Gait Index; ICC, Intraclass Correlation Coefficient; CI, 95% confidence intervals.

post stroke hemiplegia is difficult because of different statistical indices used [21]. Ceiling effects may have influenced the ICC values in a couple of ways. These effects may have decreased the amount of true variance thus leading to a reduction of the ICC values, or they may have improved the ICC values because higher agreement is observed at high scores of the scale as apparent in the Bland-Altman scatter-plots (see Figure 1).

Both the DGI scale and the BBS showed very good performance in terms of interrater reliability. This outcome is surprising since in general the BBS has a more accurate definition of items with time intervals to discriminate between the levels of each item. The BBS also requires fewer judgements concerning the 'normality' of performances. The variability of gait performances on the DGI between sessions was demonstrated by a decrease of the ICC value in the test-retest condition. In our opinion this variability was due to an inconsistency of performances rather than to inconsistency of the rater.

ICC values obtained in this study were lower than the data obtained by McConvey and Bennett [12] who reported an ICC value of 0.98 for interrater reliability. The differences in scores were likely due to different experimental procedures since McConvey used video recorded performances, thus nullifying the variance due to time effect. Because of the different statistical indices used, the comparisons with other studies were difficult [22].

The self-administered tests had similar ICC values. The ICCs for the test-retest condition were 0.92 and 0.90 for the ABC and the DHI test respectively. The ICC values obtained in this study for the ABC scale were similar to those obtained for persons with lower limb amputation [13] but higher than those obtained by Botter et al. for persons with stroke [23]. Scores seemed less stable for lower levels of perceived balance impairment as apparent from the Bland-Altman plots. No such trends were detected for the DHI test.

For the BBS and DGI scale the slight improvement in scores from the first to the second consecutive assessment was not clinically relevant. In contrast to tests based on performance, self-administered scales showed a small reduction of values from first to second assessment.

SEM refer to the variability of test scores obtained from a single person had that person been tested multiple times. The SEMs were quite low among tests with the exception of the DGI in the retest paradigm where the level of uncertainty was 4 points, an error that could be clinically relevant.

The variability of symptoms reported by persons with MS may have accounted for a small increase of variability in the clinical balance tests when compared to persons with lower limb amputation.

showed similar coefficients. Note the differences in ICC values between interrater and test-retest conditions for the DGI scale.

Bland-Altman scatter plots of the tests in the interrater and test-retest conditions are depicted in Figure 1.

A visual evaluation of Bland-Altman scatter plot of the BBS demonstrated an apparently better reliability at the higher level of scores in the interrater condition as demonstrated by a closer position of the points to the X axis; the same was true for the ABC scale in the test-retest condition. No other trends were detectable for the other tests in the two experimental conditions.

Standard errors of measurements, calculated from mean scores, are shown in Table I. For the BBS and DGI scale the SEMs were in the range of 1.5–2 points. For self-administered scales the SEMs were in the range of 7–8 points.

Discussion

The aim of this study was to assess interrater and test-retest reliability of 4 scales tapping balance disorders in MS population.

The BBS showed good performances both in terms of interrater and test-retest reliability. The ICC values proved to be satisfactory though lower than those of elderly adults [19] and traumatic brain injury persons [20]. Comparison with persons with
or stroke. This variability in symptoms may be the cause of the higher ICCs in the interrater condition than the test-retest condition.

A limitation of this study was the small number of persons enrolled and the clinical characteristics of the group that showed only mild to moderate balance impairments. The data should thus be generalized with caution to persons with different levels of impairment.

The raters were experienced neurological physiotherapists, the same level of reliability cannot be assumed with less experienced physiotherapists or other health care professions. A stronger evaluation of interrater reliability could have been gained with more raters having a wider range of expertise or with the adoption of more complex studies (e.g., D studies) to assess the reliability of an assessment carried out by different therapists in different occasions, for example two consecutive assessments rated by two different therapists [17]. The historical effects on the performance of the tests could be better evaluated by the comparison of videotaped performances.

A better description of several items and more accurate definitions of the scoring system within each item would further improve the reliability of those scales.

**Conclusion**

The results of this study support the concept that the test scores for persons suffering from multiple sclerosis are not less reliable with respect to other pathologies. The inter-rater reliability of the instruments proved to be satisfactory. Lower but acceptable results were obtained for the test-retest
paradigm suggesting that these scales are reliable tools for assessing aspects of balance function in persons suffering from MS.

We assessed the test-retest and inter-rater reliability of a couple of scales that measure balance performance and a couple of scales that measure the perception of balance, both types of scales demonstrated excellent reliability when applied to persons with MS. The comparison between balance performance, person’s balance perception and the assessment of subject’s behaviours in the activities of daily life may improve the estimation of fall risk and allow the implementation of a more tailored treatment plan.

References


