TITLE

Reliability and minimal detectable change of the mini-BESTest in adults with spinal cord injury in a rehabilitation setting

AUTHORS

Audrey Roy, PT, Johanne Higgins, OT, PhD, Sylvie Nadeau, PT, PhD

School of rehabilitation, Université de Montréal, Montréal, Canada

Center for Interdisciplinary Research in Rehabilitation of Greater Montreal (CRIR)

Institut universitaire sur la réadaptation en déficience physique de Montréal (IURDPM) du Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'Ile-de-Montréal (CCSMTL)

Corresponding author:

Sylvie Nadeau, PT, PhD
École de Réadaptation,
Faculté de médecine, Université de Montréal Pavillon du Parc, Bureau 402-18
C.P.6128 Succ. Centre-ville,
Montréal QC, Canada, H3C 3J7
Email: sylvie.nadeau@umontreal.ca

The research project was approved by the Center for Interdisciplinary Research in Rehabilitation of Greater Montreal's ethics committee.

This manuscript contains 229 words in the abstract and 3,862 words in the main text.

ABSTRACT

Background. The mini-Balance Evaluation Systems Test (mini-BESTest) is a valid tool for assessing standing balance in people with spinal cord injury (SCI). Its reliability has not yet been investigated with this population.

Objective. To assess the test-retest and inter-rater reliability of the mini-BESTest in adults with SCI in a rehabilitation setting.

Methods. Twenty-three participants admitted in a rehabilitation center following a SCI (mean age = 52.2 years, SD = 14.5; 13/23 tetraplegia; 14/23 traumatic injury) and able to stand 30 seconds without help were recruited. They were evaluated twice with the mini-BESTest to establish the test-retest reliability (interval of 1 to 2 days). One of the two sessions was video-recorded to establish the inter-rater reliability (3 physiotherapists). Intraclass correlation coefficients ($ICC_{2,1}$), weighted kappa (K_w) and Kendall's *W* were used to determine reliability of total score and individual items. Minimal detectable changes (MDC) were computed.

Results. The mini-BESTest total scores showed excellent test-retest (ICC=0.94) and inter-rater (ICC=0.96) reliability. Reliability of 50% of the individual items was acceptable to excellent (K_w and $W = 0.35 – 1.00$). The MDC of the mini-BESTest total score was 4 points.

Conclusion. The mini-BESTest is a reliable tool to assess standing balance in adults with a SCI. A minimal change of 4 points on the total scale is needed to be confident that the change is not a measurement error between two sessions or two raters.

Key Words: Reliability, mini-BESTest, Balance, Spinal Cord Injuries, Rehabilitation

INTRODUCTION

Balance control is a complex phenomenon and must be evaluated comprehensively to guide therapeutic interventions and assess improvement. A panel of experts recently reviewed all existing standing balance outcome measures and recommended the use of the Berg Balance Scale (BBS) and the mini-BESTest with adult populations (Sibley et al, 2015b). While the BBS is widely used (Bambirra, Rodrigues, Faria, and Paula, 2015; Berg, Wood-Dauphinee, Williams, and Gayton, 1989; Sibley et al, 2015b) and is a valid and reliable measure to use with the SCI population (Spinal Cord Injury Research Evidence, 2016), a ceiling effect limits its applicability for people with less balance deficits (Datta, Lorenz, and Harkema, 2012; Jørgensen et al, 2017; Lemay and Nadeau, 2010). The BBS has also been criticized for not considering more dynamic components of balance such as reactive postural control and balance during gait (Datta, Lorenz, and Harkema, 2012; Sibley et al, 2015a). Considering that approximately 38% of individuals with a SCI recover the ability to walk one year post injury (Spinal Cord Injury Model Systems, 2017), these balance components should be assessed.

The mini-BESTest is a short version of the BESTest (Balance Evaluation Systems Test) designed to comprehensively assess various components of standing balance. It is more clinically applicable than its longer version (15 minutes vs 45 minutes) (Franchignoni et al, 2015) and assesses 14 items coming from 4 of the 6 BESTest components of standing balance: anticipatory postural control (subscale I: 3 items), reactive postural control (subscale II: 3 items), sensory orientation (subscale III: 3 items) and dynamic gait (subscale IV: 5 items). Roaldsen, Wakefield, and Opheim (2015) explored the usefulness of the mini-BESTest for the rehabilitation of adults with various diagnoses, including SCI. They concluded that the mini-BESTest may help clinicians to identify the postural control components causing balance impairment and establish targeted interventions, especially for people with higher functional levels.

Regarding the psychometric properties of the mini-BESTest, Jørgensen et al. (2017) reported a good internal consistency and construct validity with chronic SCI. They also found that the score could differentiate adults with a SCI walking with/without walking aids and those having low/high concerns about falling. No ceiling effect was mentioned in any previous study (Chinsongkram et al, 2014; Chiu and Pang, 2017; Goljar et al, 2017; Hamre, Botolfsen, Tangen, and Helbostad, 2017; Jacome, Cruz, Oliveira, and Marques, 2016; Jørgensen et al, 2017; Marques et al, 2016; Roaldsen, Wakefield, and Opheim, 2015; Ross et al, 2016; Schlenstedt et al, 2015). However, the test-retest and inter-rater reliability, as well as the minimal detectable change (MDC), have not yet been established in adults with SCI in a rehabilitation setting. Thus, the aim of this study was to investigate the test-retest and the inter-rater reliability of the mini-BESTest total scores and individual items and to determine the MDC for adults undergoing intensive in-patient rehabilitation after SCI. Because the overall reliability of the mini-BESTest scores was found to be good to excellent among other groups of patients such as Parkinson’s Disease, chronic and subacute stroke, multiple sclerosis and the elderly (Dahl and Jørgensen, 2014; Di Carlo et al, 2016; Goljar et al, 2017; Hamre, Botolfsen, Tangen, and Helbostad, 2017; Marques et al, 2016; Ross, Purtill, and Coote, 2016), it was hypothesized that the test would also show good to excellent reliability for SCI adults in a rehabilitation setting.

METHODS

Study design

A prospective observational study was conducted to measure the reliability (test-retest and inter-rater) of the mini-BESTest. This article was reported based on recommendations of the guideline for reporting reliability and agreement studies (GRRAS) (Kottner et al, 2011) and on the **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments (COSMIN) (Terwee et al, 2007; Terwee et al, 2012).

Participants

Adults with a SCI admitted in a public rehabilitation center in Canada were recruited from August 2015 to September 2016. All participants had gone through spinal surgery and were in-patients undergoing functional rehabilitation aiming to maximize their independence at the time of recruitment. Participants met the following inclusion criteria: (1) aged between 18 and 75 years old, (2) sustained a traumatic or non-traumatic SCI, complete or incomplete (4) able to stand without aid for 30 seconds, (5) spoke French or English, (6) tolerated 20 minutes of evaluation with rest periods, (8) able to provide an informed consent. Participants were excluded if they had a severe neurological condition other than the SCI or a musculoskeletal or medical condition that would interfere with the measurements. They were also excluded if they had a psychiatric condition or dementia that could alter understanding of the instructions. All participants signed an informed consent form prior to the study, which was approved by the local ethics committee (CRIR-1082-0515).

Sample size

The sample size needed for this study was estimated based on an alpha level of 0.05 and a power of 0.80. The minimal acceptable level of ICC was set at 0.70 and the predicted ICC was 0.90 (confidence interval of ± 0.2). A sample of 19 participants was required to establish the test-retest (n=2 sessions) and inter-rater (n=3 raters) reliability (Walter, Eliasziw, and Donner, 1998).

Outcome measure

The mini-BESTest is a 14-item balance measure. Each task is scored on a 3-point scale (from 0 to 2) for a maximal total score of 28 and maximal sub-scores of 6 (subscales I, II, III) or 10 (subscale IV), with higher scores representing better balance (Franchignoni et al, 2010; Horak, 2018). Participants were evaluated wearing their comfortable shoes and orthoses if they needed them for safety, except for item #3 (rise up to toes) and #9 (incline – eyes closed) for which orthoses would have been restrictive. The mini-BESTest was administered according to the official instructions available on the author’s web page (Horak, 2018). Deviations from the straight line were considered as imbalance in items #10 and #11, as discussed with the author of the test. The same equipment (chair, foam surface, incline, box) was used for every participant. For French-speaking participants, the test was freely translated by the evaluator because a French version was not yet available.

Procedure

After recruitment, demographic information, as well as outcome measures routinely conducted by the rehabilitation team, were collected from the participant’s medical file. The three raters involved in the assessment were physiotherapists with at least eight years of experience in SCI rehabilitation. The three raters undertook the video training available on the author’s web page (Horak, 2018). They also participated in three sessions (total 3 hours) discussing and practicing the administration and scoring of each item to maximize standardization.

The reliability study was conducted at baseline (inclusion in the study) for half of the participants and a few days before discharge for the other half in order to have representative levels of balance impairments.

Test-retest reliability. Rater 1 evaluated all participants using the mini-BESTest twice, within a 24 to 48-hour interval. The two evaluations were made at the same period of the day to avoid influence of fatigue on the participant’s performance. Participants used the same walking device and orthoses (if needed) for both sessions. The rater was blinded to their previous ratings. A second person provided close supervision at all times for security.

Inter-rater reliability. One of the two evaluations conducted by rater 1 was recorded by two video cameras (two different angles). Raters 1, 2 and 3 looked at the videos to score each participant’s performance. Rater 1 watched the videos at least one month after the sessions to be blinded to her previous ratings. The scoring of the three raters was made independently.

Statistical analysis

Statistical analyses were performed using IBM® SPSS® Statistics version 24.0 (IBM corporation, Armonk, New York). Descriptive statistics and clinical characteristics were used to describe the sample. A Shapiro-Wilk test was used for the assessment of normality of the distributions of scores and score differences. The significance level was set to $p< .05$.

Scores from sessions 1 and 2 were used to calculate test-retest reliability and scores from the video assessments of raters 1, 2 and 3 were used to calculate inter-rater reliability.

Intraclass correlation coefficients ($ICC_{2,1}$) (2-way random analysis of variance (ANOVA), absolute agreement, single measurement) with their respective 95% confidence interval (95%CI) were computed for total scores and sub-scores. Reliability of sub-scores must however be interpreted with caution as the mini-BESTest is unidimensional (Franchignoni et al, 2015; Franchignoni et al, 2010) and these divisions, based on a postural control systems framework (Horak, Wrisley, and Frank, 2009), have not been validated. ICC values greater than 0.70 are recommended as a minimum standard for reliability (Terwee et al, 2007) and values greater than 0.80 were considered excellent (Di Carlo et al, 2016). The standard error of measurement (SEM) was calculated using each ICC computed previously. SEM represents the measurement error expressed in the same unit of measurement as the outcome measure itself:

$$\text{SEM} = \text{SD} \sqrt{}(1\text{-ICC})$$

where SD is the standard deviation of the scores obtained on the mini-BESTest from all the observations and ICC is the corresponding reliability coefficient. The minimal detectable change (MDC) was then calculated with each SEM:

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

where 1.96 is the z-value chosen. MDC_{95} represents the smallest score change, at a 95% confidence level, that can be considered a true change and not a measurement error alone (Streiner, Norman, and Cairney, 2015). The SEM and MDC_{95} were also expressed in percentage (SEM %, MDC_{95}%) of the maximal score possible on the total mini-BESTest.

Reliability of sub-scores and each individual item score on the mini-BESTest was also assessed using the quadratic weighted kappa statistic (K_w, with 95%CI) and the Kendall's coefficient of concordance (Kendall's *W,* with p-value) for test-retest and inter-rater reliability, respectively. The K_w observes the agreement between paired scores. A kappa value of 1 represents perfect agreement between the two measurements, a value of 0 indicates no more agreement than that expected by chance and a kappa value of -1 would indicate perfect disagreement between measurements (McHugh, 2012). Kendall's *W* ranks the observation from the different raters and determines how much variability there is between the average ranks. Values of K_w and Kendall's *W* were interpreted as the ICCs (Di Carlo et al, 2016; Gisev, Bell, and Chen, 2013; Landis and Koch, 1977).

The kappa statistics and Kendall's *W* cannot be produced for items having no variability (i.e. same score attributed to every participant). For these items, the percent agreement (%agreement) was calculated as described by McHugh (McHugh, 2012). This author's classification of the level of agreement was also used: 0-4% = no agreement; 4-15% =

minimal; 15-35% = weak; 35-63% = moderate; 64-81% = strong; 82-100% = almost perfect agreement (McHugh, 2012).

Finally, the Bland and Altman (B&A) plots of difference against mean with Limits of Agreement (LA) were used as a visual demonstration of the agreement between sessions and pairs of raters. T-tests were computed to detect the presence of systematic bias, in which case the mean difference (d) would be significantly different from 0. The 95%LA was calculated as follows (Giavarina, 2015):

$$95\%LA = \text{mean difference } (d) \pm 1.96{*}SD$$

where SD is the standard deviation of the differences.

RESULTS

Characteristics of participants

Thirty-two in-patients were approached for recruitment. Six refused to participate for personal reasons and two were excluded based on health problems that were among the exclusion criteria (ankle sprain and psychological condition). One participant was not re-assessed because his condition was different (needed to wear an ankle brace). Therefore, twenty-three participants with a SCI level between C2 and L5 completed the reliability study. Twelve of them were assessed at their inclusion in the study and 11, just before discharge from in-patient SCI rehabilitation. Participants' demographic characteristics at the time of assessment are presented in Table 1. Five participants needed a walking device to perform the walking items of the mini-BESTest (walker: n=4;

two canes: n=1) and two participants needed ankle-foot orthoses. Table 2 presents the average total score and sub-scores on the mini-BESTest for each session and rater.

Test-retest reliability

Relative test-retest reliability was excellent for the mini-BESTest total score (ICC = 0.94) (Table 3) and for scores on subscales I, III and IV (ICC = 0.83 – 0.93; K_w = 0.83 – 0.93). It was acceptable for the score on subscale II (ICC = 0.72; K_w = 0.71). Absolute reliability expressed with SEM is also shown in Table 3. Test-retest reliability coefficients (Table 4) were excellent for five items (K_w and % > 0.86), acceptable for three more items (K_w = 0.61 – 0.78) and below the acceptable levels for six items (K_w = 0.35 – 0.59). There was no statistically significant agreement between the two sessions for items #4 and #10 (K_w = 0.35 – 0.40, p >0.05). On the B&A plot (Figure 1A), the mean difference between total scores attributed on the two different sessions (d = 0) did not differ from zero (p = 1.00) and no heteroscedasticity was observed. Moreover, the test-retest reliability of scores obtained by the participants evaluated at baseline (ICC = 0.89) and at discharge (0.91) was similar, with 95%CI showing a largest range at baseline (0.45-0.98 vs 0.75-0.97).

Inter-rater reliability

Relative inter-rater reliability was overall excellent for total score (ICC = 0.96; Table 3) and sub-scores (ICC = 0.80 – 0.95; *W* = 0.88 – 0.99). Inter-rater reliability was also excellent for 13 of the mini-BESTest individual items (*W* = 0.83 – 1.00, p < .001; Table 4) and acceptable for item #2 (rise to toes) (*W* = 0.74). B&A plots (Figures 1B, C and D) revealed no systematic error for total score attributed by every pair of raters, with mean differences (d = ± 0.22 - 0.43) not statistically different from 0 ($p \geq 0.31$). Raters 1 and 2 showed more agreement with each other than with rater 3 with narrower limits of agreement (Figure 1).

Minimal detectable change

MDC_{95} of total scores derived from the test-retest ICC and inter-rater ICC were 3.83 vs 3.43 points respectively.

DISCUSSION

To our knowledge, this study was the first to assess reliability of the mini-BESTest in SCI adults in rehabilitation. Our hypothesis that the test-retest and inter-rater reliability of the mini-BESTest total scores would be acceptable to assess standing balance of this population is confirmed. Results are even above recommendations for clinical use (ICC > 0.90) (Streiner, Norman, and Cairney, 2015). Our ICC values (total group: 0.94 – 0.96; baseline/discharge groups: 0.89 – 0.91) are also comparable to previous studies assessing other populations (range 0.71 to 0.99) (Anson, Thompson, Ma, and Jeka, 2017; Chiu and Pang, 2017; Dahl and Jørgensen, 2014; Di Carlo et al, 2016; Goljar et al, 2017; Hamre, Botolfsen, Tangen, and Helbostad, 2017; Jacome, Cruz, Oliveira, and Marques, 2016; Jacome et al, 2017; Marques et al, 2016; Ross, Purtill, and Coote, 2016). Agreement between raters 1 and 2 was greater than with rater 3. However, every LA (-4.05 to 3.61) was comparable or inferior to estimated values from previous studies (± 3 to 6 points (Huang et al, 2016; Jacome, Cruz, Oliveira, and Marques, 2016; Jacome et al, 2017; Löfgren et al, 2014; Marques et al, 2016; Ross, Purtill, and Coote, 2016)). With a MDC of 3.43 to 3.83 points and LA around 4 points, we are confident that a score change of 4 points on the total mini-BESTest is beyond the measurement error alone and indicates a true change of the balance status for an individual. A change of 4 points corresponds to a 14.3% change on the total scale, which approaches the MDC_{95} of 10.3% (5.74 points out of 56) for the BBS with chronic incomplete SCI (Tamburella, Scivoletto, Iosa, and Molinari, 2014). The slightly lower MDC for the BBS might be explained by the different population studied (chronic SCI vs sub-acute) and type of ICC used ($ICC_{3,1}$ vs $ICC_{2,1}$). A change of 4 points is also within the range of the previously calculated MDC for the mini-BESTest in neurological populations, i.e. between 2.0 and 8.4 points (Chiu and Pang, 2017; Dahl and Jørgensen, 2014; Godi et al, 2013; Hamre, Botolfsen, Tangen, and Helbostad, 2017; Lampropoulou et al, 2018; Löfgren et al, 2014; Ross, Purtill, and Coote, 2016; Tsang, Liao, Chung, and Pang, 2013). Not considering studies with the extreme data narrows this range to MDC = 3.0 – 5.3 points, within which our MDC is still situated. The lower MDC from Dahl and Jørgensen (2014) could be explained by the methodology (ratings from video-recordings only) and the type of ICC used ($ICC_{1,1}$ and $ICC_{3,1}$) whereas the higher MDC reported by Chiu and Pang (2017) could be a consequence of the combination of the lower ICCs (0.80 – 0.81) and the higher SD probably attributable to a more heterogeneous study sample.

Four previous studies assessed reliability of individual items of the mini-BESTest (Chiu and Pang, 2017; Dahl and Jørgensen, 2014; Ross, Purtill, and Coote, 2016; Tsang, Liao, Chung, and Pang, 2013). Every study found a wide range of reliability values for items (kappa (k) and K_w = 0.21 – 1.00), which is in line with our results (K_w and *W* = 0.35 – 1.00). No item was consistently unreliable across studies, although 3 out of 4 studies reported kappa values below 0.70 for items #2 (rise to toes) and #6 (compensatory stepping correction – lateral). In the present study, item #6 was also among the less reliable items, along with the two other items from the postural response subscale (subscale II; test-retest $ICC_{2,1}$ [95%CI] = 0.72 [0.44-0.87] and inter-rater $ICC_{2,1}$ = 0.80 [0.66-0.90]). Löfgren et al. (2014) also observed lower values for this subscale. In fact, rater 1 reported that getting patients to lean their body weight correctly with support into her hands before releasing the support was difficult, which may have introduced inconsistency in how the items were performed between sessions as well as between participants. The variability from the rater's «performance» is, however, not the only source of variability because other items like «change in gait speed» and «walk with head turns – horizontal» showed even lower test-retest reliability (K_w = 0.35 and 0.44 respectively) and require no physical action from the rater.

In line with our results, item #7 (stance; eyes open, firm surface) had coefficient values over 0.70 in every previous study (k = 0.81 – 1.00). In our study, though, the higher reliability of item #7 is certainly due to its ceiling effect (everyone achieved the maximal score on this item). We believe that this ceiling effect on item #7 is a consequence of our inclusion criteria (being able to stand 30 seconds without help). This item is included in subscale III (sensory orientation) which, along with the other items in this subscale, showed the highest test-retest and inter-rater reliability coefficients (ICC, K_w and W >

0.87). Löfgren et al (2014)'s hypothesis that rating a performance based on time, as in items from subscale III, is easier than rating one based on qualitative characteristics is supported by our results. Indeed, every score based on the number of seconds holding a position showed acceptable levels of reliability. The static nature of these items could also be an explanation, because almost every item with insufficient levels of reliability involved dynamic postural control and belonged to subscales II (postural response) and IV (dynamic gait). Dahl and Jørgensen (2014) had similar findings in individuals with stroke.

The reliability of individual items of the mini-BESTest was lower for the test-retest than for the inter-rater assessments. Indeed, 50% of the items showed insufficient test-retest reliability ($K_w < 0.70$) whereas every inter-rater item score was considered reliable ($W >$ 0.70). This was also observed for the total scores and sub-scores, albeit less than for the items. The fact that video-recording was used to assess the inter-rater reliability instead of re-testing the participants is the main explanation. Our excellent video-based reliability results mean that the rating scale and the rater's clinical judgment are not responsible for the poor test-retest reliability results. The possibility of a learning effect causing the lower test-retest results is discarded due to the absence of a systematic change as shown by the B&A plots. The variability of the participants' performance is also left out with the video-recording methodology (the exact same performance of the participant is rated). Knowing that performance variability in walking balance has already been demonstrated in individuals with a chronic SCI (Day et al, 2012) and considering that our participants were in a subacute phase of recovery, we believe that a great part of our test-retest reliability results was influenced by the participants' variability in their performance. We could not exclude the possibility that our inter-rater reliability results didn’t include the variability of the rater’s instructions and actions (discussed in a previous paragraph). However, while we have to consider that the ICCs found in this study could have been slightly lower if participants were re-evaluated for inter-rater reliability, our results provide a reference value of the measurement error of the mini-BESTest for this clientele.

The choice of an adequate methodology in this study was also a challenge in terms of test-retest time-interval. Indeed, 24 to 48 hours was chosen as a time-interval to avoid changes in the participant’s condition, considering the rapid neurological recovery of some subacute SCI patients. This strategy was efficient because the B&A plots showed no systematic improvement of scores. Such a short time-interval is prone to a memory bias from the rater, even if our rater didn’t have access to the previous scoring sheets (i.e. blinded). However, we know that this possible memory bias was not important because the score means for session 2 (rater 1’s re-test) and rater 1’s inter-rater scoring (done more than one month later) were very similar (Table 2).

This study has a few other limitations. First, the results can be generalized only to people in a subacute phase of SCI. More accurately, these results are applicable to the individuals able to stand 30 seconds without help. This inclusion criterion may have prevented the lower scores (from 0 to 6/28 points) on the mini-BESTest from being tested for reliability. Moreover, while the intention was to include all types of spinal cord lesions encountered in a rehabilitation setting, no AIS A nor AIS C and no sacral lesions were represented. The fact that a French version of the mini-BESTest (Lemay, Roy, Nadeau, and Gagnon, 2018) was not available at the time of the evaluations is also among the limits of this study. Finally, the MDC is one of the several measures used to assess the responsiveness of an outcome measure. Further research would be interesting to investigate the responsiveness of the mini-BESTest in people with a SCI.

CONCLUSION

The findings in this study suggest that the mini-BESTest is a reliable outcome measure in individuals with SCI in a rehabilitation setting. A minimal change of 4 points on the total score from one session to another, or from one rater to another, is recommended to make sure that the change is not a measurement error (MDC).

IMPLICATIONS FOR CLINICAL PRACTICE

The MDC value recommended in this study is helpful for clinicians in their analysis of the change in their patients' standing balance. Furthermore, our observations on the reliability of individual items of the mini-BESTest suggest that during their training, the evaluators may need more practice in administrating dynamic items, especially those of the postural response subscale (subscale II).

ACKNOWLEDGMENTS

Audrey Roy received a scholarship from the Ordre Professionnel de la Physiothérapie du Québec, the Université de Montréal and the Center for Interdisciplinary Research in Rehabilitation of Greater Montreal (CRIR). The authors acknowledge Jean-François Lemay, Catherine Dansereau and the clinicians of the IURDPM for their collaboration throughout the project.

The authors report no conflict of interest.

Ms. Roy, Prof. Higgins, PhD and Prof. Nadeau, PhD provided concept/research design, data analysis/interpretation and writing/revision of the article. Ms. Roy provided data collection and project management.

REFERENCES

Anson E, Thompson E, Ma L, Jeka J 2017 Reliability and Fall Risk Detection for the BESTest and Mini-BESTest in Older Adults. Journal of Geriatric Physical Therapy [In Press] https://doi.org/10.1519/jpt.0000000000000123.

Bambirra C, Rodrigues MCB, Faria CDCM, Rodrigues-de-Paula F 2015 Clinical evaluation of balance in hemiparetic adults: a systematic review. Fisioterapia em Movimento 28: 187-200.

Berg K, Wood-Dauphinée S, Williams JI, Gayton D 1989 Measuring balance in the elderly: preliminary development of an instrument. Physiotherapy Canada 41: 304-311.

Chinsongkram B, Chaikeeree N, Saengsirisuwan V, Viriyatharakij N, Horak FB, Boonsinsukh R 2014 Reliability and Validity of the Balance Evaluation Systems Test (BESTest) in People With Subacute Stroke. Physical Therapy 94: 1632-1643.

Chiu AYY, Pang MYC 2017 Assessment of Psychometric Properties of Various Balance Assessment Tools in Persons With Cervical Spondylotic Myelopathy. Journal of Orthopaedic and Sports Physical Therapy 47: 673-682.

Dahl S, Jørgensen L 2014 Intra- and Inter-Rater Reliability of the Mini-Balance Evaluation Systems Test in Individuals with Stroke. International Journal of Physical Medicine & Rehabilitation 2: 177.

Datta S, Lorenz DJ, Harkema SJ 2012 Dynamic Longitudinal Evaluation of the Utility of the Berg Balance Scale in Individuals With Motor Incomplete Spinal Cord Injury. Archives of Physical Medicine and Rehabilitation 93: 1565-1573.

Day KV, Kautz SA, Wu SS, Suter SP, Behrman AL 2012 Foot placement variability as a walking balance mechanism post-spinal cord injury. Clinical Biomechanics 27: 145-150.

Di Carlo S, Bravini E, Vercelli S, Massazza G, Ferriero G 2016 The Mini-BESTest: a review of psychometric properties. International Journal of Rehabilitation Research 39: 97-105.

Franchignoni F, Godi M, Guglielmetti S, Nardone A, Giordano A 2015 Enhancing the usefulness of the Mini-BESTest for measuring dynamic balance: a Rasch validation study. European Journal of Physical and Rehabilitation Medicine 51: 429-437.

Franchignoni F, Horak F, Godi M, Nardone A, Giordano A 2010 Using psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. Journal of Rehabilitation Medicine 42: 323-331.

Giavarina D 2015 Understanding Bland Altman analysis. Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara 25: 141-151.

Gisev N, Bell JS, Chen TF 2013 Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy 9: 330-338.

Godi M, Franchignoni F, Caligari M, Giordano A, Turcato AM, Nardone A 2013 Comparison of Reliability, Validity, and Responsiveness of the Mini-BESTest and Berg Balance Scale in Patients With Balance Disorders. Physical Therapy 93: 158-167.

Goljar N, Giordano A, Schnurrer Luke Vrbanić T, Rudolf M, Banicek-Sosa I, Albensi C, Burger H, Franchignoni F 2017 Rasch validation and comparison of Slovenian, Croatian, and Italian versions of the Mini-BESTest in patients with subacute stroke. International Journal of Rehabilitation Research 40: 232-239.

Hamre C, Botolfsen P, Tangen GG, Helbostad JL 2017 Interrater and test-retest reliability and validity of the Norwegian version of the BESTest and mini-BESTest in people with increased risk of falling. BMC Geriatrics 17: 92.

Horak FB 2018 BESTest :: Home. http://bestest.us.

Horak FB, Wrisley DM, Frank J 2009 The Balance Evaluation Systems Test (BESTest) to Differentiate Balance Deficits. Physical Therapy 89: 484-498.

Huang MH, Miller K, Smith K, Fredrickson K, Shilling T 2016 Reliability, Validity, and Minimal Detectable Change of Balance Evaluation Systems Test and Its Short

Versions in Older Cancer Survivors: A Pilot Study. Journal of Geriatric Physical Therapy 39: 58-63.

Jácome C, Cruz J, Oliveira A, Marques A 2016 Validity, Reliability, and Ability to Identify Fall Status of the Berg Balance Scale, BESTest, Mini-BESTest, and Brief-BESTest in Patients With COPD. Physical Therapy 96: 1807-1815.

Jácome C, Flores I, Martins F, Castro C, McPhee CC, Shepherd E, Demain S, Figueiredo D, Marques A 2017 Validity, reliability and minimal detectable change of the balance evaluation systems test (BESTest), mini-BESTest and brief-BESTest in patients with end-stage renal disease. Disability and Rehabilitation [In Press] https://doi.org/10.1080/09638288.2017.1375034.

Jørgensen V, Opheim A, Halvarsson A, Franzén E, Roaldsen KS 2017 Comparison of the Berg Balance Scale and the Mini-BESTest for Assessing Balance in Ambulatory People With Spinal Cord Injury: Validation Study. Physical Therapy 97: 677-687.

Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A, Johansen M, Jones L, Krassioukov A, Mulcahey MJ, et al. 2011 International standards for neurological classification of spinal cord injury (Revised 2011). Journal of Spinal Cord Medicine 34: 535-546.

Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL 2011 Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. International Journal of Nursing Studies 48: 661-671.

Lampropoulou SI, Billis E, Gedikoglou IA, Michailidou C, Nowicky AV, Skrinou D, Michailidi F, Chandrinou D, Meligoni M 2018 Reliability, validity and minimal detectable change of the Mini-BESTest in Greek participants with chronic stroke. Physiotherapy Theory and Practice [In Press] https://doi.org/10.1080/09593985.2018.1441931: 1-12.

Landis JR, Koch GG 1977 The Measurement of Observer Agreement for Categorical Data. Biometrics 33: 159-174.

Lemay JF, Nadeau S 2010 Standing balance assessment in ASIA D paraplegic and tetraplegic participants: concurrent validity of the Berg Balance Scale. Spinal Cord 48: 245-250.

Lemay JF, Roy A, Nadeau S, Gagnon DH 2018 French version of the Mini BESTest: A translation and transcultural adaptation study incorporating a reliability analysis for individuals with sensorimotor impairments undergoing functional rehabilitation. Annals of Physical and Rehabilitation Medicine [In Press] https://doi.org/10.1016/j.rehab.2018.12.001.

Löfgren N, Lenholm E, Conradsson D, Ståhle A, Franzén E 2014 The Mini-BESTest - a clinically reproducible tool for balance evaluations in mild to moderate Parkinson's disease? BMC Neurology 14: 235.

Marques A, Almeida S, Carvalho J, Cruz J, Oliveira A, Jácome C 2016 Reliability, Validity, and Ability to Identify Fall Status of the Balance Evaluation Systems Test, Mini-Balance Evaluation Systems Test, and Brief-Balance Evaluation Systems Test in Older People Living in the Community. Archives of Physical Medicine and Rehabilitation 97: 2166-2173.e2161.

McHugh ML 2012 Interrater reliability: the kappa statistic. Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara 22: 276-282.

Roaldsen KS, Wakefield E, Opheim A 2015 Pragmatic Evaluation of Aspects Concerning Validity and Feasibility of the Mini Balance Evaluation System Test in a Specialized Rehabilitation Hospital. International Journal of Physical Therapy & Rehabilitation 1: 104-109.

Ross E, Purtill H, Coote S 2016 Inter-rater reliability of mini balance evaluation system test in ambulatory people with multiple sclerosis. International Journal of Therapy and Rehabilitation 23: 583-589.

Ross E, Purtill H, Uszynski M, Hayes S, Casey B, Browne C, Coote S 2016 Cohort Study Comparing the Berg Balance Scale and the Mini-BESTest in People Who Have Multiple Sclerosis and Are Ambulatory. Physical Therapy 96: 1448-1455.

Schlenstedt C, Brombacher S, Hartwigsen G, Weisser B, Möller B, Deuschl G 2015 Comparing the Fullerton Advanced Balance Scale With the Mini-BESTest and Berg Balance Scale to Assess Postural Control in Patients With Parkinson Disease. Archives of Physical Medicine and Rehabilitation 96: 218-225.

Sibley KM, Beauchamp MK, Van Ooteghem K, Straus SE, Jaglal SB 2015a Using the Systems Framework for Postural Control to Analyze the Components of Balance Evaluated in Standardized Balance Measures: A Scoping Review. Archives of Physical Medicine and Rehabilitation 96: 122-132.e129.

Sibley KM, Howe T, Lamb SE, Lord SR, Maki BE, Rose DJ, Scott V, Stathokostas L, Straus SE, Jaglal SB 2015b Recommendations for a Core Outcome Set for Measuring Standing Balance in Adult Populations: A Consensus-Based Approach. PloS One 10.

Spinal Cord Injury Model Systems 2017 2017 Annual Report - Public Version. National spinal cord injury statistical center. Birmingham, Alabama. https://www.nscisc.uab.edu/Public/2017%20Annual%20Report%20-%20Complete%20Public%20Version.pdf.

Spinal Cord Injury Research Evidence 2016 Home - Spinal Cord Injury Research Evidence. Spinal Cord Injury Research Evidence. Vancouver, British Columbia. https://scireproject.com.

Streiner DL, Norman GR, Cairney J 2015 Health Measurement Scales : A practical guide to their development and use. Oxford, United Kingdom: Oxford University Press.

Tamburella F, Scivoletto G, Iosa M, Molinari M 2014 Reliability, validity, and effectiveness of center of pressure parameters in assessing stabilometric platform in subjects with incomplete spinal cord injury: a serial cross-sectional study. Journal of Neuroengineering and Rehabilitation 11: 86.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC 2007 Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 60: 34-42.

Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC 2012 Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 21: 651-657.

Tsang CS, Liao LR, Chung RC, Pang MY 2013 Psychometric Properties of the Mini-Balance Evaluation Systems Test (Mini-BESTest) in Community-Dwelling Individuals With Chronic Stroke. Physical Therapy 93: 1102-1115.

Walter SD, Eliasziw M, Donner A 1998 SAMPLE SIZE AND OPTIMAL DESIGNS FOR RELIABILITY STUDIES. Statistics in Medicine 17: 101-110.

TABLES

	Mean	SD	Range
Age (years)	55.2	14.5	24.3 - 73.3
Time post-surgery (days)	49.3	28.6	8 - 135
Body mass index (kg m^{-2})	24.6	3.6	18.7 - 32.5
LEMS (/50)	44.9	4.3	35 - 50
10MWT max (m s^{-1}) (n=21)	1.16	0.48	0.44 - 1.92
6MWT (m)	245.0*	-	175 - 440**
BBS (/56)	50.0*	-	39 - 54**
SCIM-III (/100)	83.0*	-	50 - 95**
	N	%	
Male/Female	17/6	74.0/26.0	
Language French/English	20/3	87.0/13.0	
Tetraplegia/Paraplegia	13/10	56.5/43.5	
AIS B/D	1/22	4.3/95.7	
T/NT	14/9	60.9/39.1	

Table 1. Demographic characteristics at the time of reliability assessments (n=23). LEMS = Lower extremity motor score, 10MWT max = 10-meter walking test at maximal speed**,** 6MWT = 6-minute walk test**,** BBS = Berg Balance Scale**,** SCIM-III = Spinal Cord Independence Measure version 3**,** AIS = American Spinal Injury Association Impairment Scale (B = Sensory incomplete spinal cord injury, D = Motor incomplete spinal cord injury; Kirshblum et al, 2011), T/NT = traumatic/ non-traumatic injury.
* Median.
** Interquartile range.

	Session 1 mean (SD) [range]	**Session 2 mean (SD) [range]**	**Rater 1 mean (SD) [range]**	**Rater 2 mean (SD) [range]**	**Rater 3 mean (SD) [range]**
Total mini-BESTest (/28)	17.5 (5.9) [7-28]	17.9 (5.7) [8-28]	17.6 (5.9) [7-28]	17.4 (5.8) [7-28]	17.8 (6.1) [7-28]
I Anticipatory (/6)	3.8 (1.4) [1-6]	3.9 (1.5) [1-6]	3.8 (1.4) [1-6]	3.7 (1.3) [1-6]	3.6 (1.5) [1-6]
II Postural response (/6)	3.3 (1.9) [0-6]	3.3 (1.8) [0-6]	3.3 (2.2) [0-6]	3.4 (2.1) [0-6]	3.6 (1.5) [1-6]
III Sensory orientation (/6)	5.0 [5-6]*	5.0 [5-6]*	5.0 [5-6]*	5.0 [5-6]*	5.0 [5-6]*
IV Dynamic gait (/10)	5.2 (3.0) [0-10]	5.4 (2.7) [1-10]	5.2 (2.9) [0-10]	5.1 (2.8) [0-10]	5.3 (3.0) [1-10]

Table 2. Mini-BESTest means of the total scores and sub-scores (n=23).
Mean total scores and sub-scores on the mini-BESTest obtained on each session and by each rater. SD = standard deviation. Maximal score possible in each section and in the total mini-BESTest are specified following the respective title.
*Median [interquartile range]

	$ICC_{2,1}$ (95%CI)	**SEM**	**SEM %**	**MDC_{95}**	**MDC_{95}%**
Test-retest reliability	0.94 (0.87-0.97)	1.40	5.0	3.83	13.7
Inter-rater reliability	0.96 (0.91-0.98)	1.24	4.4	3.43	12.3

Table 3. Reliability and minimal detectable change of the mini-BESTest total score (n=23).
$ICC_{2,1}$ = intraclass correlation coefficient, SEM = standard error of measurement, SEM% = standard error of measurement expressed in percentage of maximal score, MDC_{95} = minimal detectable change with 95% confidence level and MDC_{95}%= minimal detectable change with 95% confidence level expressed in percentage of maximal score.

Items	**Sessions 1 vs 2 (Test-retest)**	**Raters 1, 2 and 3 (Inter-rater)**
	K_w (95%CI)	*W* (p)
1. Sit to stand	0.54 (0.18-0.90)	0.92 (<.001)
2. Rise to toes	0.74 (0.55-0.93)	0.74 (=.001)
3. Stand on one leg	0.86 (0.72-1.00)	0.98 (<.001)
4. Compensatory stepping correction - forward	0.40 (-0.44-0.83)	0.92 (<.001)
5. Compensatory stepping correction - backward	0.61 (0.37-0.85)	0.96 (<.001)
6. Compensatory stepping correction - lateral	0.59 (0.27-0.91)	0.93 (<.001)
7. Stance; eyes open, firm surface	100.0%*	100.0%*
8. Stance; eyes closed, foam surface	0.87 (0.71-1.00)	0.98 (<.001)
9. Incline eyes closed	1.00 (1.00-1.00)	1.00 (<.001)
10. Change in speed	0.35 (-0.55-0.75)	0.86 (<.001)
11. Walk with head turns - horizontal	0.44 (0.05-0.84)	0.83 (<.001)
12. Walk with pivot turns	0.78 (0.60-0.97)	0.87 (<.001)
13. Step over obstacles	0.90 (0.80-1.00)	0.93 (<.001)
14. Timed up & go with dual task	0.50 (0.19-0.80)	0.84 (<.001)

Table 4. Agreement between sessions and raters' scores on the mini-BESTest (n=23).
K_w = quadratic weighted Kappa statistic, 95% CI = 95% confidence interval, *W* = Kendall's coefficient of concordance, p = statistical significance (level of significance set at 0.05).
*Agreement expressed in percent agreement.

FIGURE LEGENDS

Figure 1.

Bland and Altman plots of agreement between sessions (A) and between raters (B, C, D) for the mini-BESTest, where d = mean difference between scores (with standard deviation), ±1.96*SD = 95% limits of agreement, SD = standard deviation.

FIGURES

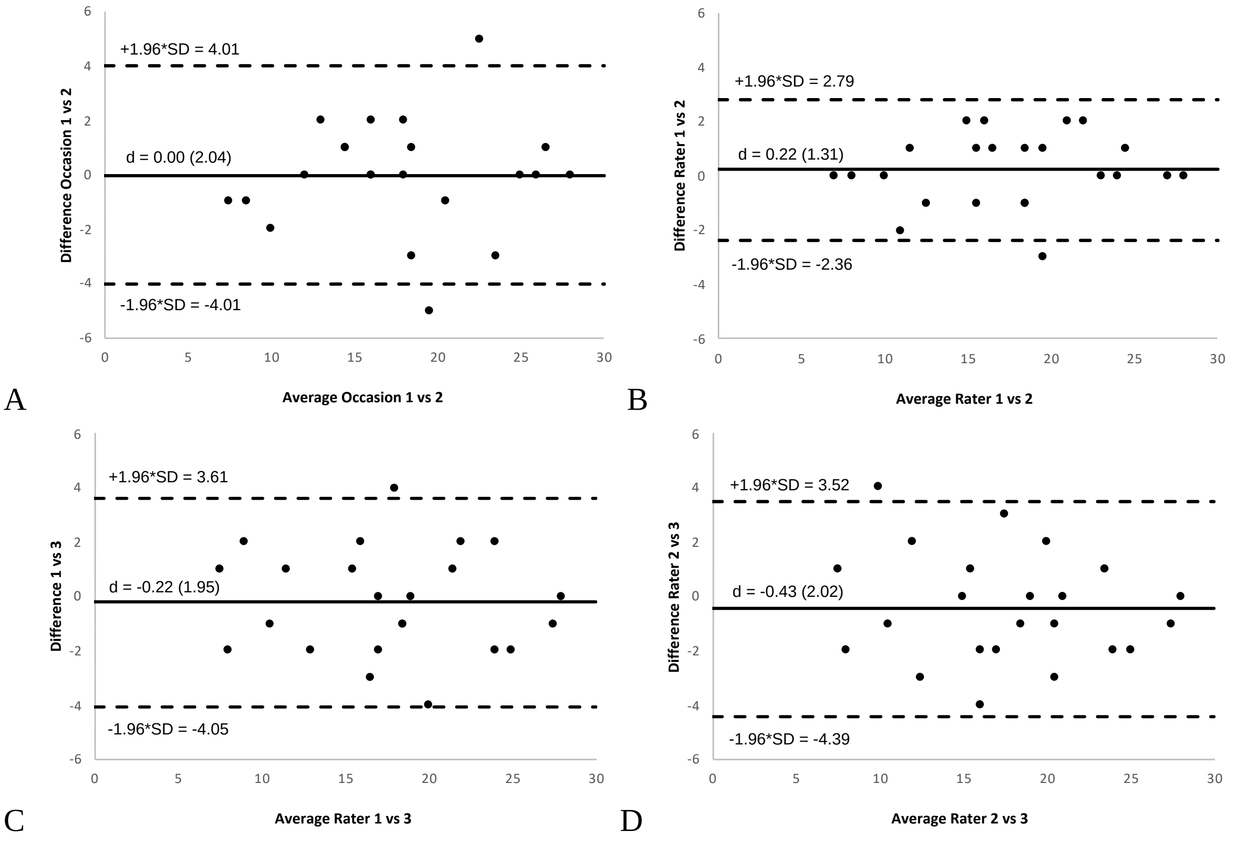

Figure 1.
Bland and Altman plots of agreement between sessions (A) and between raters (B, C, D) for the mini-BESTest, where d = mean difference between scores (with standard deviation), ±1.96*SD = 95% limits of agreement, SD = standard deviation.