The Test-Retest Reliability and Minimal Detectable Change of the Balance Error Scoring System

Daniel James Amin1, Jaclyn Coleman2, and Lee C. Herrington3

1. Department of Sport, University Centre Doncaster, Doncaster, South Yorkshire DN57SZ, United Kingdom
2. Physiotherapy Department, Heywood, Middleton & Rochdale Primary Care Trust, Rochdale, Lancashire OL161XU, United Kingdom
3. College of Health and Social Care, University of Salford, Salford, Lancashire M66PU, United Kingdom

Abstract: Context: Balance is an important aspect of sports performance, which can be impaired following MHI (minor head injury). The BESS (balance error scoring system) is a tool commonly used for assessment of post-MHI recovery. Objective: This study aimed to ascertain test-retest reliability and minimal detectable change of the BESS in order to ensure that it had appropriate clinical levels of reliability and sensitivity. Design: Test-retest reliability. Setting: University indoor sport facility. Participants: N = 36 sub-elite athletes. Interventions: The BESS was undertaken by the participants on two separate testing sessions, 1 week apart. Main Outcome Measures: Test-retest reliability and minimal detectable change. Results: Strong test-retest reliability of the total BESS score was identified (ICC (intraclass correlation coefficients) = 0.784) with a MDC (minimal detectable change) of 6-10 error points, dependent on confidence interval used. Conclusions: The strong reliability score and the relatively low MDC when using conservative measures would recommend the use of BESS in post-MHI assessment of young academy-level athletes. However, more research is required to further determine BESS MDC in both sub-elite and elite groups of athletic performers so as to provide clinicians with an accurate guideline regarding whether an athlete is presenting with post-MHI balance impairment.

Key words: Concussion, balance, reliability, sensitivity.

1. Introduction

1.1 Balance and Minor Head Injury

“Balance” is defined as the “ability to maintain a position within the limits of stability or base of support” [1]. The factors affecting balance involve somatosensory, visual and vestibular systems as well as musculoskeletal coordination, joint range of motion and strength [1, 2]. MHI (minor head injury) is traumatically induced physiological disruption of brain function, as manifested by at least one of the following: any period of loss of consciousness; any loss of memory for events immediately before or after the accident; any alteration in mental state at the time of the accident (e.g., feeling dazed, disoriented or confused); and focal neurological deficit(s) that may or may not be transient [3]. MHI is known to have detrimental effects on sensorimotor functioning of this type, through the impairment of the neuromusculoskeletal and/or sensory systems, and has, therefore, long been associated with loss of balance [3, 4].

Post-MHI, it is important to monitor the athlete’s recovery and essential to make a correct return-to-play decision, as it has been reported that, in contact sport, players who sustain an MHI are 3 times more likely to sustain another during the same season [5]. In recent years, various balance assessments have become more ubiquitous when evaluating sport-related MHI [6, 7].

The BESS (Balance Error Scoring System) [8] is one such assessment, which is said to provide an objective method of assessing static balance after MHI [9]. Information obtained from this clinical balance tool can be used to assist clinicians as part of the

Corresponding author: Daniel James Amin, M.Sc., research field: balance and pedagogy. E-mail: dan.amin@don.ac.uk.
process in making post-MHI, return-to-play decisions. Baseline scores are usually ascertained at the beginning of the season when the athlete is healthy, then re-administered post-MHI to measure the athlete’s recovery [10]. If an athlete’s performance does not improve on subsequent test administrations, then clinicians should assume that an enduring abnormality may be present. BESS scores have been shown to correlate well with force plate stability measures in post-MHI athletes ($r = 0.78$-$0.96$) [7]. However, Valovich et al. [9] found that administration of multiple trials of the BESS, spaced two days apart, resulted in a learning effect, with the number of errors decreasing with each consecutive trial. This learning effect can last up to two weeks [11] and should alert clinicians to be cautious when administering the BESS post-MHI to assess recovery in an athlete, as this could mask the actual improvements due to neurological recovery [10].

1.2 Reliability and Sensitivity of Clinical Assessment Tools

A method of assessment needs to be psychometrically robust, in terms of reliability and sensitivity, for it is to be used with confidence in the clinical setting. The test-retest reliability of an assessment tool is important to know, as it aims to ensure that a tool is reliable and consistent when measuring a clinical trait of a participant. Inconsistencies can be caused by tester error, variations in the testing procedure, or changes in the participant’s physical or mental condition [12]. The test-retest reliability is of particular importance when considering that there could be a learning effect present as in the case of the BESS. Previous studies have found the BESS to be classified as “moderate” in terms of test-retest reliability, when considering young athletes (aged 9-14; intraclass correlations ICC$(2, 1) = 0.71$) [13] and young adults (mean age: 20.4; $G$ (generalizability) = 0.64) [14].

Identifying a change score is another important psychometric measure as it highlights the sensitivity of a tool and whether a change in score is clinically relevant or simply attributed to measurement error [15]. Various methods have been utilized to calculate change scores, including MCID (minimal clinically important difference), which is determined subjectively or objectively and is the smallest meaningful change, as judged by the patient or experts in the field [16]. RCI (reliability change index) is also used to calculate change scores and assesses intra-individual differences over time [6], and has been investigated in the BESS [13] and other neuropsychological measures of concussion [17]. MDC (minimal detectable change), which is the smallest meaningful change in terms of a quantifiable objective measure of subjective performance, however, is the most commonly used in the literature [16]. Once the MDC is determined on a particular test for a given population, therapists can interpret whether the change score for their patient is at or above this minimal level of detectable change. Scores are due to patient improvement on the test rather than measurement error [18].

In the case of the BESS, MDC would need to be known to ensure that post-MHI rehabilitation is the reason for any changes in BESS score and, conversely, to ascertain whether balance has been affected by the MHI; however, to the knowledge of the authors, no previously published investigations have reported the MDC of the BESS.

The objectives of this study, therefore, are to ascertain the level of test-retest reliability of the BESS in order to ensure that it has clinically acceptable reliability and also to identify the MDC, so that a measure of true change can be established.

2. Methods

2.1 Participants

Thirty six healthy participants (25 female, 11 male; age = 19.66 ± 2.60 years; height = 170.66 ± 8.25 cm).
Inclusion criteria required that each participant was performing at sub-elite level within their respective sports (19 trampolinists, 11 soccer players, six dancers). The following exclusion criteria were applied: current musculoskeletal or orthopaedic condition affecting lower limb function; any head injury during the previous six months; previous lower limb surgery; any neurological, visual or vestibular condition affecting balance.

The procedures followed protocol with the ethical standards of the responsible institutional review board and was granted approval, in the spirit of the Helsinki declaration.

2.2 Procedures

The BESS protocol was developed by researchers and clinicians at the University of North Carolina’s Sports Medicine Research Laboratory, Chapel Hill, NC 27599-8700 and is readily accessible online. It shall be referred to as the “protocol” in these procedures.

According to the protocol, participants performed 3 barefoot stances: double-leg, single-leg, and tandem. Each stance lasted 20 seconds, whereby the participants attempted to maintain postural stability, and was performed on both a firm surface and on a 47 × 39 × 6.5 cm³ block of medium-density foam (Power Systems Airex Balance Pad 81000, Knoxville, TN.), with eyes closed and hands on the iliac crests.

The BESS has set criteria for scoring errors in the form of error points and were counted by the investigator. Each error was scored as one error point and included: opening the eyes; stepping, stumbling, and falling out of the test position; lifting the hands off the iliac crests; lifting the toes or heels; moving the leg into more than 30° of flexion or abduction; and remaining out of the test position for more than 5 seconds [13]. Multiple errors occurring simultaneously were scored as one error point, whereas an inability to maintain the test position for more than 5 seconds was scored as 10 error points. The total number of error points for each of the six testing conditions were calculated and summed for the participant’s total BESS error score. A maximum of 10 error points was attributed to each testing condition, thus a maximum total error score of 60 was attributed to the BESS. Participants were instructed on how to undertake each test condition via the script within the protocol and undertook each test 3 times, with a mean being used for further analyses [19].

In order to ensure that the research design focussed on assessing test-retest reliability, participants were re-tested only seven days later in order to try and ensure that a minimal amount of physiological change took place [12] and to mimic the re-test time-frame when assessing athlete’s recovering post-MHI. Any participant who had suffered an injury during this time was removed from the investigation completely (three in total). The investigator had four years clinical experience and had received previous training on BESS administration, within three months of undertaking the investigation.

2.3 Statistical Analyses

All statistical analyses were conducted using SPSS for Windows Version 16.0. Alpha levels were set at 0.05 for all tests.

Independent samples t-tests were carried out to determine any differences between gender and between sports, and intraclass correlation coefficients (ICC(2, 1)) were carried out for analysis of test-retest reliability of the BESS. ICC(2, 1) was specifically used as, in effect, only one judge, taken at random from a larger population of judges, scored the errors [20]. The data, therefore, were analysed within SPSS using a two-way random, single measures analysis. Absolute agreement was used as the study looked to identify the test-retest reliability when considering each individual participant [21]. Test-retest reliability strength based on ICC values was classified accordingly [22]: “weak” (< 0.50), “moderate” (0.50-0.75), “strong” (> 0.75), “ideal” (> 0.90). MDC was calculated to
establish random error and, therefore, a “true” difference in subsequent scores, using the following formula [16]: 

\[
MDC = \text{z-score}_{\text{level of confidence}} \times SD_{\text{baseline}} \times \sqrt{2 \left[ 1 - r_{\text{test-retest}} \right]}
\]

- z-score represents the confidence interval from a normal distribution;
- SD_{\text{baseline}} is the standard deviation at baseline;
- \( r \) is the test-retest reliability coefficient;
- The multiplier of \( \sqrt{2} \) was used to account for the additional uncertainty introduced by using difference scores from measurements at 2 points in time [16].

Some researchers [16, 23] suggested using a confidence interval of 90% due to its use being more common in the literature; however, a confidence interval of 95% increases the precision of score estimation and is also commonly known in the literature as the SDD (smallest detectable difference) [16]. More conservative confidence intervals of 70% and 80% have also been used previously when considering post-concussive balance assessment [13], as such; MDC’s using 70%, 80%, 90% and 95% confidence intervals were measured here.

3. Results

Independent samples t-tests confirmed no significant differences in total BESS score, on either testing session, between genders or between sport-types (see Table 1). As such, ICC and MDC analyses were undertaken on the whole subject group.

Table 2 summarises the key findings of the statistical analyses for test-retest reliability, measurement error and 95% confidence intervals for the various test conditions. \( ICC_{(2,1)} \) values for the individual test conditions and the total error ranged from 0 to 0.823, \( ICC_{(2,1)} \) values were significant in the following conditions—total BESS error score, single-leg firm and foam, tandem leg firm and double leg foam. Table 3 highlights the MDC values for the BESS conditions at various levels of confidence. These values ranged from 5.527 to 9.207 raw error points for total BESS error score.

### Table 1  Significant differences between gender and sport for various test conditions.

<table>
<thead>
<tr>
<th>Test groups</th>
<th>Test condition(s)</th>
<th>T-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male vs. female</td>
<td>Retest session (tandem, firm)</td>
<td>-2.059</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>Retest session (double-leg, foam)</td>
<td>-2.224</td>
<td>0.034</td>
</tr>
<tr>
<td>Dancers vs. trampolinists</td>
<td>Initial test session (tandem, firm)</td>
<td>-3.122</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Retest session (single-leg, firm)</td>
<td>2.964</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Retest session (double-leg, foam)</td>
<td>-2.781</td>
<td>0.012</td>
</tr>
<tr>
<td>Footballers vs. trampolinists</td>
<td>Initial test session (single-leg, foam)</td>
<td>2.187</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>Retest session (tandem, firm)</td>
<td>-2.336</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>Retest session (single-leg, foam)</td>
<td>-2.407</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>Retest session (double-leg, foam)</td>
<td>2.252</td>
<td>0.039</td>
</tr>
</tbody>
</table>

### Table 2  Values for mean, standard deviation and intraclass correlation coefficient for the various BESS conditions and total error scores.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Initial test session</th>
<th>Retest session</th>
<th>ICC(_{(2,1)})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Error Points raw</td>
<td>SD</td>
<td>Mean error points raw</td>
</tr>
<tr>
<td>Double-leg, firm surface</td>
<td>0</td>
<td>0.13</td>
<td>0.34</td>
</tr>
<tr>
<td>Single-leg, firm surface</td>
<td>2.06</td>
<td>2.06</td>
<td>1.39</td>
</tr>
<tr>
<td>Tandem-stance, firm surface</td>
<td>1.75</td>
<td>1.66</td>
<td>0.88</td>
</tr>
<tr>
<td>Double-leg, foam surface</td>
<td>0.81</td>
<td>1.81</td>
<td>0.06</td>
</tr>
<tr>
<td>Single-leg, foam surface</td>
<td>6.08</td>
<td>2.77</td>
<td>6.63</td>
</tr>
<tr>
<td>Tandem-stance, foam surface</td>
<td>4.08</td>
<td>2.61</td>
<td>4.88</td>
</tr>
<tr>
<td>Total error score</td>
<td>14.78</td>
<td>7.69</td>
<td>14.44</td>
</tr>
</tbody>
</table>

\( ICC_{(2,1)} \): intraclass correlation coefficient; SD: standard deviation. *P < 0.05.
The Test-Retest Reliability and Minimal Detectable Change of the Balance Error Scoring System

Table 3  Minimal detectable change for each BESS condition at varying levels of confidence.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MDC (minimal detectable change) based on level of confidence, raw score (test unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MDC$_{95}$</td>
</tr>
<tr>
<td>Double-leg, firm surface</td>
<td>0</td>
</tr>
<tr>
<td>Single-leg, firm surface</td>
<td>4.102 (5)</td>
</tr>
<tr>
<td>Tandem-stance, firm surface</td>
<td>2.780 (3)</td>
</tr>
<tr>
<td>Double-leg, foam surface</td>
<td>2.111 (3)</td>
</tr>
<tr>
<td>Single-leg, foam surface</td>
<td>4.510 (5)</td>
</tr>
<tr>
<td>Tandem-stance, foam surface</td>
<td>6.344 (7)</td>
</tr>
<tr>
<td>Total error score</td>
<td>9.907 (10)</td>
</tr>
</tbody>
</table>

4. Discussion

4.1 Principle findings

In order for an assessment tool to be recommended for clinical practice, it must be classified psychometrically as having “good” reliability, validity and sensitivity [22]. The results show that the BESS has varying levels of test-retest reliability when considering each test condition (0-0.823). However, the test-retest reliability for total error score (0.784), which is the score used within the clinical environment, is classed as being “strong” [22]. These results compare favourably with the literature, which have found moderate test-retest reliability amongst children ($ICC_{2,1} = 0.71$) [13] and adults (generalizability = 0.64) [14]. The latter study found that if an average of two test scores were taken at two time points, then the reliability increased to a more clinically acceptable $G = 0.82$-$0.85$, when sexes were examined independently. Direct comparisons are difficult here, however, as generalizability is a different statistical methodology to the intraclass correlation coefficient methodology used in this study [22]. But the concept of taking an average of multiple tests at each session, as was administered in this investigation and advocated elsewhere [19], seems to be beneficial for ensuring higher test-retest values. The study with 9-14 years old subjects [13] did not administer this multiple test approach, so may have had an improved test-retest value if it did, and achieve a “strong” clinical outcome, as the value is only 0.4 from the statistical indicators. The strong test-retest reliability results found in this investigation coupled with that already seen in the literature suggests that the BESS test is a consistent clinical tool. The “weak” reliability values associated with the single-leg, firm surface and tandem-stance, foam surface conditions may suggest a favourable learning effect [11] present. However, this is still not being seen consistently enough in the literature to warrant an alteration to the protocol at this time.

The MDC for total error score, which highlights the amount of change in error points required for the change to be clinically relevant, ranges from 6 to 10 based on level of confidence. Based on the largest confidence interval MDC$_{95}$ (SDD), for a “true” change in total error score to have taken place, a change of > 9 error points is required. In the clinical setting, this would suggest that a swing of > 9 from baseline would be required before a clinician could consider whether neurological impairment has occurred; conversely, if the test was initially administered post-MHI to monitor recovery, the athlete would need to show a reduction in score of > 9 error points in order to show true signs of neurological improvement. Using more conservative confidence intervals (MDC$_{70}$, MDC$_{80}$ and MDC$_{90}$) as advocated in under-18 athletes [24], would mean a smaller change in error points required to infer a true change in score (6, 7 and 9, respectively). When comparing the more conservative MDC$_{70}$, MDC$_{80}$ and MDC$_{90}$ of total BESS error score to the RCI’s of similar confidence intervals identified in the study by Valovich McLeod et al. [13], we can see some similarity with only 1 unit difference at each
of these confidence levels. The SDD for total BESS error score of 10 found in this investigation can be considered high since an average difference of seven errors (range: 6-9 errors) from baseline score or control group has been seen in post-concussive subjects [25, 26]. Using SDD in this instance has the potential implication that an athlete, who has undergone MHI and suffered noticeable balance impairment, may not have this change registered as true if undertaking the BESS.

4.2 Considerations for the BESS

The reliability of the BESS may be enhanced further when considering any possible test-design limitations.

When a lone administrator administers the test, in situ, it is difficult to score all errors that are actually occurring. For instance, some errors occur in different planes (e.g., 30° flexion of the leg is in the sagittal plane, whereas, 30° abduction of the leg is in the frontal plane—where does the administrator stand to ensure they capture both of these potential errors?). Some errors occur at various height levels (e.g., an administrator may be looking for the eyes opening, yet miss the heel raising). It is also difficult to accurately determine the time frame of 5 seconds, which has implication upon various error scores. This may be rectified by using video replay to confirm all the errors that have taken place, as seen previously [27]. Previous studies into other dynamic balance assessment tools have found that using video replay can increase reliability scores [28, 29]. Also, multiple errors occurring simultaneously are suggested to be scored as only one error, despite the multiple postural deviations occurring. Once again, video replay could potentially measure more accurately the number of errors occurring, which, when bearing in mind the nature of the test in terms of assessing recovery from MHI, maybe worthwhile despite the obvious time costs.

It would be interesting to determine whether test-retest reliability and MDC, which is calculated using test-retest correlation coefficients, is enhanced due to the experience of the administrators. Considering what the BESS is assessing, it is vital that all test administrators can accurately score the errors consistently. It may, therefore, require a clinician to prove their proficiency prior to administering this test with their athletes. Previous literature in the field of post-concussive assessment, with particular reference to BESS, suggested that training is encouraged to establish consistency among multiple raters and that this training should be highlighted [19, 24]. However, the actual training period required for someone to be able to administer the test is not referred to in the literature. In the case of this investigation, the experience of the test administrator was three months and a strong test-retest reliability was achieved, however, the mean total error score was 14.06, which is high when considering the range of means in healthy subjects identified during a recent systematic review of the BESS (8.4-14.1 total error score) [19]. It may have been that the test administrator in this study was being overly liberal when scoring errors, which is partly confirmed when comparing the mean total error score of the soccer players in this study (14.73 ± 6.03) with previous literature assessing total error score in soccer players (12.5 ± 5.16) [30]. However, it must be stated that the gender of the subjects were not comparable between these two studies. As such, it is still uncertain as to whether tester expertise will have an effect on reliability of the BESS test.

4.3 Limitations and Future Research Possibilities

The limitations of this study included the fact that the test administrator was possibly not experienced enough as a BESS tester; however, as has been discussed prior, there are still no set recommendations for training period prior to clinical administration. Also, the participants were all healthy athletes, which does not infer to the basis of the BESS test. However, this use of healthy participants is in alignment with
other similar BESS test studies [1, 9, 27]. It is now important to further determine the MDC of the BESS, using various subject groups, so as to provide clinicians with an accurate guideline regarding whether an athlete is presenting with post-MHI balance impairment, however, it must be noted that MDC should not be the single factor in making return-to-play decisions, post-MHI; instead, this true change must be taken into consideration along with a variety of clinical assessment methods both subjective and objective [13].

4.4 Clinical Implications

There is currently no recognised gold standard measure for balance activity in patients with neurological conditions [10], however, this study adds to the body of literature in terms of demonstrating that the test-retest reliability of the BESS is strong when using healthy, young adult academy athletes. It also advocates the protocol recommendations of using multiple administrations per session in order to enhance the reliability of the scores recorded [19]. However, there is still room for identifying a minimum trainer experience to see whether this factor has in influence on improving test-retest reliability of the BESS, and thus providing a more accurate MDC. Identifying that $MDC_{95}$ is greater than the differences seen in the literature for BESS error score pre- and post-concussion, may suggest that the BESS is not sensitive enough to carry out clinical testing on MHI, however, it has to be noted that the literature has recommended more conservative assessment measures when assessing in this field [24]. So it may be that using an MDC with a lower confidence interval is preferable. In which case, $MDC_{70}$ and $MDC_{80}$ would be sensitive enough to identify these noted changes. It is also worth pointing out that there is still a paucity in the literature in terms of using concussed athletes as subjects, as such, the average change in score seen pre- and post-MHI thus far may not be a true representation of the changes that take place.

References

The Test-Retest Reliability and Minimal Detectable Change of the Balance Error Scoring System


