Intratester and Intertester Reliability During the Star Excursion Balance Tests

Jay Hertel, S. John Miller, and Craig R. Denegar

Objective: To estimate intratester and intertester reliability and learning effects during the Star Excursion Balance Tests (SEBTs).

Setting: A university athletic training research laboratory.

Subjects: Sixteen healthy volunteers with no history of balance disorders or significant lower extremity joint pathology.

Measurements: Length of excursion was measured manually for each trial.

Results: ICCs for intratester reliability were .78-.96 on day 1 and .82-.96 on day 2. ICCs for intertester reliability were .35-.84 on day 1 and .81-.93 on day 2. Significant learning effects were identified for 4 of the 8 tests.

Conclusions: Estimates of intratester and intertester reliability were high, but adequate practice trials should be performed before taking baseline measures.

Key Words: dynamic balance, postural control, functional performance, motor learning


Quantification of dynamic balance has been advocated for use as a clinical measure of functional performance following injury; however, the vast majority of dynamic balance protocols reported in the rehabilitation literature were developed for the pediatric and geriatric populations or for those recovering from cerebrovascular accidents or traumatic brain injuries. Few noninstrumented dynamic balance protocols have been reported that effectively challenge the postural control systems of well-conditioned, physically active individuals recovering from lower extremity injuries. One such test battery is the Star Excursion Balance Tests (SEBTs).

The SEBTs are a series of 8 unilateral balance tests that incorporate a single-leg stance of 1 leg with a maximum targeted reach of the free leg. The stance leg operates in the closed kinetic chain with coupled motion at the ankle, knee, and hip joints as the opposite leg reaches in the specified direction. As the targeted reach is performed with the foot, the postural control system is challenged as the body’s center of mass is moved in rela-
tion to its base of support. Postural control is mediated by the visual, vestibular, and somatosensory systems. Adequate neuromuscular control of the muscles of the stance leg is paramount to increasing the length of excursion of the reach leg. Thus, optimal performance of these tests can only be obtained if there are no restrictions to range of motion or neuromuscular control at the ankle, knee, and hip joints.

Although these tests hold promise as an effective noninstrumented, clinically applicable dynamic balance test battery, the reliability and validity of these tests have not been extensively reported. One previous investigation has examined the intratester reliability and learning effects of 4 of the SEBTs. To date, there have been no reports of the intratester reliability or learning effects of all 8 of the SEBTs and no investigation of the intertester reliability of the SEBTs. Therefore, the purposes of this investigation were to (1) estimate the intratester reliability of the SEBTs, (2) estimate the intertester reliability of the SEBTs, and (3) identify learning effects associated with repeated trials of the SEBTs.

Methods

Subjects

Sixteen recreationally active, healthy young adults (8 men, 8 women, age = 21.3 ± 1.3 years, height = 171.2 ± 6.7 cm, mass = 70.3 ± 10.0 kg) volunteered to serve as subjects. They each completed an injury history questionnaire, and all were free of cerebral concussions, vestibular disorders, and lower extremity pathology for the 6 months prior to testing. Before participating, all subjects read and signed an informed consent form that was approved by the university’s institutional review board.

Performance of the SEBTs

The SEBTs were performed with the subject standing at the center of a grid laid on the floor with 8 lines extending at 45° increments from the center of the grid (see Figure 1). The subject maintained a single-leg stance on the stance leg while reaching with the opposite leg to touch as far as possible along the chosen line. The subject touched the farthest point possible on the line with the most distal part of their reach foot. The reach foot touched the farthest point on the line as lightly as possible so that the reach leg did not provide considerable support in maintaining upright posture. The subject then returned to a bilateral stance while maintaining equilibrium. The examiner marked the point touched along the line and then manually measured the distance from the center of the grid to the touch point with a tape measure.

Trials were discarded and repeated if the examiner felt that the reach foot was used to provide considerable support when touching the ground,
if the subject lifted the stance foot from the center of the grid, or if the subject lost his or her equilibrium at any point in the trial.

The 8 lines were named anterolateral (AL), anterior (A), anteromedial (AM), medial (M), posteromedial (PM), posterior (P), posterolateral (PL), and lateral (L), according to the direction of excursion in relation to the stance leg; thus the labeling of the grid was different for the right and left legs (see Figure 2).

**Protocol**

Subjects performed 2 bouts of the 8 directions of excursions on each leg on each of 2 days. Subjects performed 1 bout while being assessed by examiner 1 and performed the other bout while being assessed by examiner 2 on each day. The order of examiners was counterbalanced so that a different examiner assessed each subject first on days 1 and 2. Eight of the subjects began all bouts by performing the right-stance-leg tests first, and the other 8 subjects began all bouts by first performing the left-stance-leg tests.

![Figure 1](image)

*Figure 1* Subject performing the anterolateral excursion test.
Each subject performed 1 warm-up trial in each of the 8 directions on each leg before recording began. All subjects began by performing 3 trials in the anterolateral direction. Trials were separated by 15 seconds of rest. The order of testing for all subjects was AL, A, AM, M, PM, P, PL, and L. After performing all excursions on the initial stance leg, the same protocol was repeated with the contralateral leg serving as the stance leg. The protocol was then repeated with examiner 2 taking the measurements. Subjects returned for identical follow-up testing 1 week after day 1 testing.

**Statistical Analysis**

For each of the 8 directions of excursion, intratester reliability was estimated for each examiner by calculating the intraclass correlation coefficients (ICC$_{3,1}$) and SEMs for the 3 trials in each bout on both days.$^{13}$ Intertester reliability was estimated for the 8 directions on days 1 and 2 by calculating ICCs and SEMs for the 6 trials of each test performed each day.

Means for the 3 trials in each bout (trials 1–3, trials 4–6, trials 7–9, and trials 10–12) were calculated for each of the 8 excursions. A within-subjects repeated-measures analysis of variance (ANOVA) was performed, with the independent variable being order (trials 1–3, trials 4–6, . . .) to identify any learning effects that might have occurred with repetitive trials of the SEBTs. The 8 levels of the dependent variable were the 8 directions of excursion. Tukey post hoc analysis was used to identify specific differences. Paired $t$ tests with Bonferroni corrections were also used to compare the excursion lengths of right- and left-leg trials in each direction. The level of significance was preset at .05 for all analyses.

**Results**

ICCs and SEMs for intratester and intertester reliability for all 8 tests are shown in Tables 1 and 2, respectively. The ICCs for intratester reliability
for examiner 1 ranged from .78 to .96 on day 1 and from .85 to .96 on day 2, and those for examiner 2 ranged from .83 to .94 on day 1 and from .82 to .96 on day 2. The ICCs for intertester reliability ranged from .35 to .84 on day 1 and from .81 to .93 on day 2.

A significant learning effect was found using the within-subjects ANOVA ($F = 1.98; df = 90, 48; P = .003$). Post hoc analysis revealed significant learning effects for the L, PM, and PL directions on both legs and for the P direction on the right leg only ($P < .05$; see Figures 3–6). Means and SDs of trials 7–9 are shown in Table 3 as a pilot report of normative performance data.
Table 2  Intertester Reliability Estimates for Days 1 and 2*

<table>
<thead>
<tr>
<th>Direction</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anterolateral</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.80 (3.52)</td>
<td>.93 (2.27)</td>
</tr>
<tr>
<td>left</td>
<td>.78 (3.66)</td>
<td>.86 (2.96)</td>
</tr>
<tr>
<td>Anterior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.78 (3.99)</td>
<td>.88 (2.87)</td>
</tr>
<tr>
<td>left</td>
<td>.76 (3.95)</td>
<td>.89 (2.75)</td>
</tr>
<tr>
<td>Anteromedial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.84 (3.40)</td>
<td>.87 (2.90)</td>
</tr>
<tr>
<td>left</td>
<td>.76 (3.92)</td>
<td>.89 (2.78)</td>
</tr>
<tr>
<td>Medial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.69 (4.41)</td>
<td>.88 (3.06)</td>
</tr>
<tr>
<td>left</td>
<td>.69 (4.15)</td>
<td>.93 (2.45)</td>
</tr>
<tr>
<td>Posteromedial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.78 (3.57)</td>
<td>.89 (2.80)</td>
</tr>
<tr>
<td>left</td>
<td>.80 (3.08)</td>
<td>.93 (2.33)</td>
</tr>
<tr>
<td>Posterolateral</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.66 (4.72)</td>
<td>.88 (3.76)</td>
</tr>
<tr>
<td>left</td>
<td>.76 (3.68)</td>
<td>.91 (2.69)</td>
</tr>
<tr>
<td>Lateral</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>.61 (4.96)</td>
<td>.86 (3.87)</td>
</tr>
<tr>
<td>left</td>
<td>.58 (4.16)</td>
<td>.88 (2.83)</td>
</tr>
</tbody>
</table>

*Intraclass correlation coefficients are listed first; SEMs are in cm and in parentheses.

Comments

Intratester Reliability

These results indicate that the estimates of both intratester and intertester reliability of the SEBTs were relatively high. The ICCs for intratester reliability in our study ranged from .78 to .96. Estimates of intratester reliability indicate the consistency of measures with repeated trials as assessed by the same examiner. Previously, Kinzey and Armstrong\textsuperscript{12} reported ICCs for intratester reliability estimates of the 4 diagonal excursion tests (AM, AL, PM, and PL) to range from .67 to .87 in a group of 20 healthy young adults. Our reliability estimates were slightly higher than theirs were.\textsuperscript{12} Likewise, the SEMs for the intratester reliability estimates in our study (1.77–3.38 cm) were lower than those previously reported (3.43–4.78 cm).
Figure 3  Results of the lateral excursion tests during right-leg stance.
*Significantly greater than mean of trials 1–3.
#Significantly greater than mean of trials.

Figure 4  Results of the posterolateral excursion tests during right-leg stance.
*Significantly greater than trials 1–3.

Our intratester reliability estimates might be higher because we included uniplanar excursions (A, M, P, L) in addition to diagonal excursions—these tests were less challenging to the postural control system, and thus performance was less variable—or because our subjects performed more trials overall than did the subjects in Kinzey and Armstrong’s study,12 and variability of performance thus diminished with practice effects.

The estimates of intratester reliability on the right and left limbs were very similar in our study. This finding is important clinically because it suggests that within-subject side-to-side comparisons can be used without concern that reliability of measures is normally higher on 1 side than the other.
The estimates of intratester reliability rate comparably with those of other noninstrumented functional balance tests. The clinical test most similar to the SEBTs that has been previously reported is the functional reach (FR) test. The FR test is performed by having the subject maintain a bilateral stance while trying to reach as far forward with 1 hand as possible without moving the position of the feet. The maximum distance of forward reach is measured manually by the examiner. Duncan et al. reported very high intratester reliability (ICC = .92) of the FR test in a group of elderly men, and Light et al. reported an ICC of .98 in a group of 30 male and female elderly subjects. Donahoe et al. reported ICCs from .83 to .92 in a group of healthy children ranging in age from 5 to 15 years performing the FR test,
Table 3  Means and SDs for Trials 7–9 (Cm), During Which Performances Reached a Plateau

<table>
<thead>
<tr>
<th>Excursion direction</th>
<th>Left-leg stance</th>
<th>Right-leg stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterolateral</td>
<td>80.4 (7.7)</td>
<td>80.3 (7.8)</td>
</tr>
<tr>
<td>Anterior</td>
<td>72.4 (7.5)</td>
<td>73.8 (8.2)</td>
</tr>
<tr>
<td>Anteromedial</td>
<td>70.0 (8.6)</td>
<td>69.5 (8.0)</td>
</tr>
<tr>
<td>Medial</td>
<td>90.8 (9.1)</td>
<td>91.6 (8.8)</td>
</tr>
<tr>
<td>Posteromedial</td>
<td>89.4 (8.5)</td>
<td>91.8 (7.5)</td>
</tr>
<tr>
<td>Posterior</td>
<td>83.5 (8.1)</td>
<td>86.5 (10.7)</td>
</tr>
<tr>
<td>Posterolateral</td>
<td>83.9 (7.8)</td>
<td>83.7 (10.9)</td>
</tr>
<tr>
<td>Lateral</td>
<td>78.5 (6.7)</td>
<td>76.1 (8.2)</td>
</tr>
</tbody>
</table>

and Niznik et al.\textsuperscript{12} reported ICCs from .89 to .97 in a group of children with lower extremity spasticity performing the same test.

Several other measures of functional or dynamic balance have been reported to be reliable when used to assess balance impairment in the elderly. These include Tinnetti’s Performance Oriented Mobility Assessment (POMA),\textsuperscript{9} the Frailty and Injuries: Cooperative Studies of Intervention Techniques (FICSIT),\textsuperscript{8} and Berg’s Balance Scale.\textsuperscript{3}

The Clinical Test of Sensory Interaction and Balance (CTSIB) is a noninstrumented test of static balance that has been shown to have very high intratester reliability (Pearson’s $r = .99$).\textsuperscript{14} This test uses differing support surfaces (hard floor, foam) and visual conditions (eyes open, eyes closed, visual conflict dome) to challenge the vestibular, visual, and somatosensory systems in maintaining upright stance.

**Intertester Reliability**

Estimates of intertester reliability indicate the consistency of measures between examiners. No previous studies of the intertester reliability of the SEBTs have been reported. The estimates of intertester reliability in our study were better on day 2 than on day 1. On day 1, ICCs ranged from .35 to .84, whereas they ranged from .81 to .93 on day 2. SEMs ranged from 3.08 to 4.96 cm on day 1 and from 2.27 to 3.87 cm on day 2. The escalation in ICCs on day 2 was likely a result of the plateau seen in performance scores on the second day of testing and the reduction of variability in individual performance of the tests. Although these changes seem to be plausibly explained by reduced variability of the subjects’ performance on the tests, improvement in intertester reliability also could be a result of the evaluators becoming more consistent in their measurement techniques.
Previous reports have estimated the intertester reliability for the FR test to be ICC = .98 in groups of both healthy children\(^1\) and elderly subjects.\(^5,7\) The FR test is the only previously reported test of dynamic balance that required examiners to measure the distance of a limb excursion by the subject. ICCs of greater than .95 were reported for the intertester reliability of the scoring of the scaled variables used in the Berg Balance Scale in a group of patients recovering from strokes.\(^3\) The estimates of intertester reliability of timed variables were reported to be Pearson’s \(r = .99\) for the CTSIB in young healthy adults\(^14\) and to range from ICC = .91 to .99 in a battery of balance and gait tasks administered to a group of 24 patients recovering from hip fracture.\(^6\) Cipriany-Dacko et al\(^15\) reported no marked differences between the assessment scores of the BPOMA tests administered to the same group of elderly subjects by novice and experienced examiners.

Riemann et al\(^16\) estimated the intertester reliability of the Balance Error Scoring System (BESS) to range from ICC = .78 to .96 in assessing static balance in a group of healthy collegiate athletes. The BESS is a battery of static balance tests that incorporates different stances (unilateral, bilateral, tandem) and support surfaces (hard floor, foam) to challenge a subject’s maintenance of postural control. In scoring the BESS, the clinician observes the subject for gross corrections in the maintenance of postural control, such as stepping forward or lifting the hands off of the iliac crests.\(^16\)

### Learning Effects

Our results suggest that significant learning effects are present with repetitive trials of 4 of the 8 excursion directions. Significant differences were seen with repetitive trials of the tests requiring L, PL, P, and PM excursions. Interestingly, these 4 excursions did not allow the subject to easily visualize the target line. The significant improvements in these directions might be related to improvements stemming from practice effects related to coping with impaired visual feedback. Balance is regulated by sensory integration of the vestibular, visual, and somatosensory systems. Tests that do not allow visual feedback of the target for the excursion leg might challenge the postural control system to rely on a greater degree of sensory information from the vestibular and somatosensory systems. Significant improvements with repetitive trials in these 4 tests might occur as the postural control system accommodates to these reaching tasks in the absence of visual feedback.

It should be noted that the longest excursions occurred during trials 7–9 for all directions. This indicates that at least 6 practice trials in each direction should be performed before recording baseline data for clinical or experimental purposes. Using the Spearman–Brown prophecy statistic, Kinzey and Armstrong\(^12\) estimated that 6 practice sessions of 5 trials in each direction would be necessary to raise intratester reliability to levels above .86 and thus recommended using extensive practice trials before recording
baseline data. We concur that practice trials are necessary but believe that our analysis demonstrates that 6 practice trials in each direction would be sufficient to account for learning effects with repetitive trials in the different directions.

Subjects in our study were not given instructions that placed limitations on the strategies employed in order to reach their maximum targeted excursions. Learning effects might be related to individual subjects becoming proficient at the particular strategy that allows them to perform optimally on these tasks. Additionally, variability in performance might have occurred because there was no way to ensure that subjects were performing at maximal effort on each trial. In the future, estimation of perceived exertion during these functional tests might help control for performance variability related to subjects' alterations in effort throughout testing.

Validity

Although these tests appear to be reliable, this study did not attempt to validate them as a clinical measure of lower extremity functional capacity. The SEBTs are a unique set of tests that require maintaining a single-leg stance while attempting to reach as far as possible with the opposite leg in a given direction. The stance leg is free to undergo coupled movements at the ankle, knee, and hip, and trunk motion is also permitted as the subject attempts to attain the maximum reach with the excursion leg. As a subject attempts to attain maximum reach, his or her center of gravity is typically lowered as the closed kinetic chain motions of ankle dorsiflexion, knee flexion, and hip flexion occur on the stance leg. Depending on the direction of excursion, varying degrees of hip rotation and trunk motions also are employed.

In order for maximum excursion to occur, 2 basic biomechanical principles must be adhered to. First, the center of gravity must be adequately located over the base of support of the stance leg. For example, with an anterior excursion, the reach leg is undergoing hip flexion and displacing the body's center of gravity anteriorly in relation to the base of support. To counteract this, trunk extension is required in order to maintain the center of gravity within the limits of the base of support. The second principle that must be adhered to is that of eccentric and isometric neuromuscular control of the joints of the stance leg. Thus, execution of the SEBTs requires not only sensory integration of the vestibular, visual, and somatosensory systems but also adequate neuromuscular control in order to maximize performance. We feel that the addition of neuromuscular control to the paradigm for optimal performance makes the SEBTs an excellent method to assess dynamic balance and functional capacity of the lower extremity.

The gold standard in the assessment of static balance is the assessment of postural control variables on an instrumented forceplate. However, the relationship between the assessment of dynamic balance and the assess-
Reliability of the Star Excursion Balance Tests

The reliability of the Star Excursion Balance Tests (SEBTs) for assessing static balance is not currently understood. Two studies have been reported in which significant correlation was found between simultaneous tester subjective evaluation and objective postural control variables on a forceplate in the assessment of static balance tasks. We are not aware of any studies that have attempted to correlate performance on tests of dynamic balance with forceplate measures.

This investigation was the first in a series of experiments we are undertaking to develop the SEBTs as a useful clinical and experimental method to assess dynamic balance. There remains a need to validate the usefulness of the SEBTs in clinical populations. It is currently unknown whether the SEBTs can identify performance decrements between injured and uninjured limbs. If the tests are significantly different between injured and uninjured limbs there is also a need to document improvement of test scores with rehabilitation, as has been done with the FR test. It is also unknown whether certain populations have superior baseline performance scores on the SEBTs, as is the case with professional dancers having higher baseline scores than non-dancers on the CTSIB.

Finally, the relationship between impaired performance on the SEBTs and injury prediction should also be explored.

In conclusion, the SEBTs have high levels of both intratester and intertester reliability and appear to be a functional test of dynamic balance. Significant learning effects also appear to be present, and thus at least 6 practice trials in each direction are recommended before baseline data for clinical or experimental data are recorded. Although the SEBTs are reliable, the validity of these tests in assessing dynamic balance has yet to be established.

References


