# Generalizability of the Limits of Stability Test in the Evaluation of Dynamic Balance Among Older Adults

*Sean Clark, MS, Debra J. Rose, PhD, Koichiro Fujimoto, PhD*

ABSTRACT. Clark S, Rose DJ, Fujimoto K. Generalizability of the limits of stability test in the evaluation of dynamic balance among older adults. Arch Phys Med Rehabil 1997;78:1078-84.

**Objective:** Reliability of platform posturography tests is essential for the identification and treatment of balance-related disorders. The purposes of this study were to establish the reliability of the limits of stability (LOS) test and to determine the relative variance contributions from identified sources of measurement error.

**Design:** Generalizability theory was used to calculate (1) variance estimates and percentage of variation for the sources of measurement error, and (2) generalizability coefficients. Random effects repeated measures analysis of variance (RM ANOVA) was used to assess consistency of measurements across both days and targets.

**Participants:** Thirty-eight community-dwelling older adults with no recent history of falls.

**Main Outcome Measures:** Outcome measures derived from the LOS tests included movement velocity (MV), maximum center of gravity (COG) excursion (ME), end point COG excursion (EE), and directional control (DC).

**Results:** Estimated generalizability coefficients for 2 and 3 days of testing ranged from .69 to .91. Relative contributions of the day facet were minimal. The RM ANOVA results indicated that for three of the movement variables, no significant differences in scores were observed across days.

**Conclusions:** The 75% and 100% LOS tests are reliable tests of dynamic balance when administered to healthy older adults with no recent history of falls. Dynamic balance measures were generally consistent across multiple evaluations.



OVER THE COURSE of the past two decades, a number of commercially produced diagnostic instruments (eg, isokinetic dynamometers, computerized posturography, kinetic treadmills) have been introduced into the field of rehabilitation. These diagnostic instruments are now used on a daily basis by clinicians desiring more objective and quantifiable evaluations of patient status. Although these sophisticated diagnostic instruments are capable of providing the practitioner with more articulate measures of patient status, the degree to which these instruments produce reliable measures of performance has been largely ignored.

In the area of balance assessment, the development of computerized posturography offers the practitioner a means for conducting more comprehensive, objective evaluations of the multiple dimensions of balance (eg, the integration and organization of sensory inputs used to maintain upright balance and volitional and reactive balance control in dynamic environments). Although the information derived from computerized posturography tests can improve a clinician's ability to identify and provide treatment for balance-related disorders, the usefulness of these measurements ultimately depends on their reliability.[1,2] Unfortunately, few studies have been conducted for the purpose of establishing the reliability of the performance scores derived from this equipment.[3-6]

Reliability is defined as the degree or extent to which a measurement system is capable of providing consistent test scores that are free from error across multiple testing sessions or between multiple raters.[7-9] A measurement is considered to be reliable when similar or consistent test results are obtained from multiple evaluations of an individual. Although some degree of inconsistency in the resulting test scores is to be expected, the magnitude of these differences should not be statistically significant.[10] Despite attempts to standardize test procedures and protocols, inherent variability or inconsistencies in the observed test scores across repeated evaluations may still exist. For example, even if the same clinician evaluates a patient's performance on a particular balance test at the same time of day, under similar test conditions, with the same set of instructions, different test scores may still result from each testing session.

Observed differences or variability in test scores that arise from repeated evaluations constitute one source of measurement error.[8,10] Measurement errors also arise from multiple sources within a given measurement protocol.[11-13] Examples of these error sources may include manual test coding errors, the use of multiple testers, misunderstood test instructions by the patient, and inaccurate calibration of the equipment. Knowledge of the various sources of measurement error is therefore important for optimizing the reliability of a measurement protocol.[13-15] In fact, a clinician's ability to first identify the various sources of measurement error and then control for or eliminate these error sources will significantly influence the reliability of the measurement protocol or test instrument used.

Although estimates of the various sources of measurement error provide insight for optimizing measurement design, questions regarding the degree of inconsistency in the observed performance scores still exist. When administering patient evaluations across multiple sessions, practitioners generally observe differences or inconsistencies in a patient's performance scores that range from small to large. Practitioners often assume, however, that the magnitude of such inconsistencies is small and, therefore, not statistically or clinically significant. If this assumption is erroneous, an inaccurate assessment of the patient's performance and an inappropriate diagnosis may result.[16]

An accepted method by which to assess the degree of inconsistency in repeated performance scores is to perform an $F$ test based on the repeated measures analysis of variance (ANOVA) results.[8,10,16] A significant $F$ test indicates statistical differences in the test scores across the multiple evaluations. In such conditions where statistically significant differences exist, reliability

From the Motor Behavior Laboratory, Department of Exercise and Sport Science, Oregon State University, Corvallis, OR.

Submitted for publication September 27, 1996. Accepted in revised from February 9, 1997.

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the authors or upon any organization with which the authors are associated.

Reprint requests to Debra J. Rose, PhD, Motor Behavior Laboratory, 24 Women's Building, Oregon State University, Corvallis, OR 97331-6802.



0003-9993/97/7810-4202$3.00/0

has not been adequately established.[10,16] Subsequently, the practitioner must then assess whether the statistically significant changes in the test scores are of clinical importance.

To date, the few studies that have been conducted for the purpose of estimating the reliability of various balance-related diagnostic tests have employed various intraclass correlation coefficient (ICC) models[3-6] (see Shrout and Fleiss[17] for a review of the different ICC models). Reliability of the limits of stability (LOS) test, a test of dynamic balance available on the Balance Master,* has recently been assessed by Henderson and colleagues.[5] These investigators estimated the test-retest reliability of the LOS test when performed on two occasions 1 week apart in a sample of both young and old healthy adults ($n$ = 46). Movement variables representing the individual's ability to shift the center of gravity (COG) quickly and accurately through space (ie, movement time and path sway) demonstrated moderate to high test-retest reliability.

Liston and colleagues[4] have also examined the reliability of dynamic balance tests available on the Balance Master.* A randomized version of the LOS test was administered to a sample of hemiparetic patients on three separate occasions at 1-week intervals. Once again, the movement variables, movement time and path sway, were found to be strongly reliable (ICC(2,1) = .88 and .84, respectively) for this measurement design.

As clinicians continue to use computerized posturography for the assessment of postural control, establishing the reliability of the various tests available on these systems is critical. Reliable measures are essential when attempting to identify individuals who are at risk of falling, as well as for establishing appropriate baseline measures necessary for assessing the effectiveness of a balance intervention program. Although a small number of studies have examined the test-retest reliability of computerized posturography, the number of test sessions needed to establish consistent test scores prior to diagnostic classification or the introduction of an intervention has not been established.[3,6] Additionally, sources of measurement error associated with the various test protocols have not been adequately identified. Knowledge of these error sources, as well as their relative contributions to the total measurement error, would enhance the practical application of these studies. Specifically, clinical researchers with a knowledge of previous findings could better control or eliminate potential sources of measurement error, thereby strengthening the reliability of their own measurement protocol.

The primary purpose of this study was to estimate the reliability of the LOS test conducted at 75% and 100% of the theoretical limits of stability in a group of healthy older adults. Secondary purposes were to estimate the relative contribution of various sources of measurement error associated with the measurement design and to determine the consistency of dynamic balance measures across three test days.

## METHODS

### Subjects

Thirty-eight community-dwelling healthy older adults (21 women and 17 men) volunteered to participate in this investigation. Subjects ranged in age from 51 to 84 years ($\bar{x}$ = 67.5yrs, SD = 8.4). Subjects' height ranged from 1.47m to 1.85m ($\bar{x}$ = 1.7m, SD = .04m). All subjects completed a medical questionnaire before the first testing session. No subject reported a history of progressive neurological, cardiovascular, or musculoskeletal disease. Additionally, no subjects were currently taking any medications known to adversely affect balance or compensate for balance-related problems. Also, all subjects reported normal and/or corrected vision (eg, glasses, contacts) and had no difficulty viewing the video screen. All subjects ambulated independently and had not sustained a fall while performing daily activities during the previous 2 years. All subjects provided written informed consent before participation in the study.

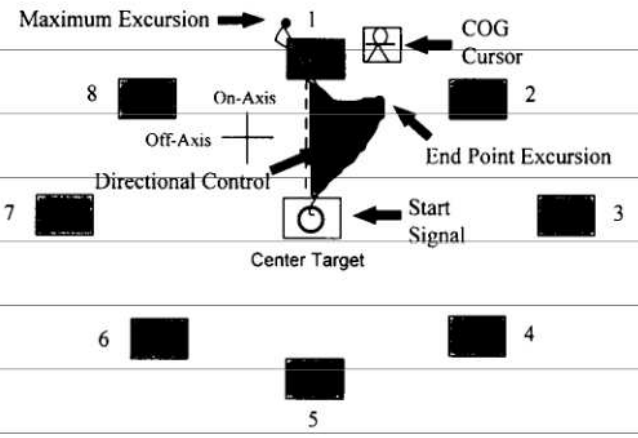


**Fig 1. Limits of Stability test: target set-up, including start signal, COG cursor, and dynamic balance measures (maximum excursion, end point excursion, and directional control).**

### Instrumentation

The Pro Balance Master* system with software version 5.0b was used in the present study. The Pro Balance Master system is comprised of two 9" × 18" dual force plates connected at the midline of the anteroposterior axis by a pin joint. Each force plate is mounted on a pair of symmetrically positioned force transducers. The four transducers measure vertical ground reaction forces (VGRF), which form the basis of subsequent calculations of center of pressure (COP) and COG sway angles.[18] All test data were acquired and then stored on a 486 PC.

### Procedures

After a brief period of familiarization (5 to 10 minutes) with the COG visual biofeedback utilized during the testing procedures, subjects completed the dynamic LOS test performed at 75% and 100% of the subject's maximum theoretical stability limits. The two LOS tests were administered in a single testing session on three consecutive days. A 3-minute rest interval was provided between each LOS test during each test session. The LOS test (performed at 75% LOS) is a standard test used to assess multiple indices of dynamic balance performance. All testing sessions were conducted at the same time of day on each of the three test days to control for potential diurnal effects. Each test was conducted with subjects positioned in a standardized foot position as recommended by the manufacturers of the equipment.[18] A reference grid superimposed on the forceplate ensured accurate placement of the feet on each testing occasion.

The dynamic LOS test is designed to assess an individual's ability to volitionally move the COG to eight predetermined positions in space. These positions are represented by square visual targets located on a video screen positioned at eye level and directly in front of the individual being tested. Figure 1 illustrates the on-screen test target set-up. Subject height (ie, subject's predicted COG height) and previously determined maximum COG sway angles were used to determine the appropriate LOS values represented by the on-screen visual targets.[18]

Subjects were required to lean away from the midline in the direction of each of the eight on-screen targets without stepping or moving their feet from the standardized foot position. Foot position was carefully monitored during each test and the feet

were repositioned following a loss of balance or any other slight foot shift during leaning. Subjects were also directed to keep their arms by their sides at all times during the test. The concurrent visual biofeedback COG cursor appeared on the video screen throughout the test. Subjects were instructed to move the COG cursor as quickly and accurately as possible towards the highlighted target as soon as a visual signal, in the form of a circle, moved from the center starting target. Subjects were then required to maintain the COG cursor inside the highlighted target until the visual signal returned to the center target. If the subject was unable to reach the target, he or she was instructed to lean as far as possible in the direction of the target without losing balance. The following dependent variables were calculated for each of the eight targets associated with the test: movement velocity (MV); maximum COG excursion (ME); endpoint COG excursion (EE); and directional control (DC).

### Dependent Variables

Each of the dependent variables provided specific information regarding the subject's LOS test performance. MV indicated the speed of the COG displacement, in degrees per second, during the first sustained movement excursion toward the test target. This variable provided an indication of how quickly the subject was able to initially move the COG through the region of stability. The degree to which the COG was controlled during the first movement excursion was expressed as DC. The DC value was derived from a comparison of the amount of on-axis movement of the COG relative to the off-axis COG movement and was expressed as a percentage of the total on-axis movement. The more direct the COG movement was toward the test target (ie, along a straight line towards the target) the smaller the DC ratio. EE indicated the furthest on-axis distance the COG reached by the end of the first sustained COG excursion towards the test target. ME indicated the furthest on-axis distance the COG traveled from the center target during the entire trial for each target. Both EE and ME are expressed as a percentage of the test target distance. Figure 1 illustrates EE, ME, and DC.

### Data Analysis

Data analyses were conducted to examine the reliability of each of the dependent variables measured in each of the two LOS tests conducted (ie, 75% and 100% of maximum LOS). The GENOVA computer program (Version 2.2) was used to analyze all data.[19] Analyses of both measurement consistency and generalizability were conducted using a fully crossed $38 \times 3 \times 8$ (subjects $\times$ day $\times$ target) random effects RM ANOVA design.

***Measurement consistency.*** The consistency of the LOS variables across the eight targets and three testing days was determined by performing tests of statistical significance for the calculated quasi-$F$ ratios based on mean squares from the RM ANOVA output.[20] For these analyses, the alpha level of significance was adjusted to $p < .01$ to minimize the inflation of type I error. Tukey post hoc comparisons of means were conducted if significant differences across days were evident. The alpha level for post hoc comparisons was also set at $p < .01$.

***Generalizability theory.*** Generalizability theory (G-theory) permits the researcher to identify and estimate the relative contribution of numerous sources of measurement error within a single model.[8,12,13] Also, G-theory distinguishes between two types of studies: a G (generalizability) study and a D (decision) study.[8,13] In the G study, the researcher identifies the various measurement conditions, or "facets," that contribute to the variability in the subjects' scores. A facet is defined as a "set of similar conditions of measurement" (eg, trials, days, weeks).[15] After the facets have been identified, ANOVA techniques are used to quantify the amount of variance attributable to differences between subjects (ie, universe score variance), the identified facets and their interactions, and the random measurement error.[8,12] Variance estimates are then used to calculate the percentage of variance associated with each of the sources of measurement error. This percentage of variance enables the researcher to identify the various sources of measurement error contributing to the variability in the subjects' scores.[8,12-14]

The subsequent D study provides the data used to make definitive decisions about the measurement protocol.[8,15] The D study yields the generalizability coefficient (G) that reflects the reliability or dependability of the observed scores across the facet(s) or measurement condition(s) included in the study.[15] Consequently, a number of different G coefficients can be obtained in any given study. The number of estimated G coefficients is dependent on the number of facets in the design and the various universes of generalization.[8,14]

The advantage of using G-theory to investigate test reliability is that this method allows for the examination of (1) the degree to which the identified sources of variance contribute to the total amount of measurement error, and (2) the generalizability or reliability of the test across the facets of interest. Findings from both the G and D studies provide a means by which to optimize the reliability associated with a particular measurement design. At a practical level, the inherent features of G-theory allow the clinician to minimize measurement error by first identifying all the possible sources of error (eg, multiple testers, test instructions, etc) and then controlling for or eliminating them.

***Application of generalizability theory.*** To perform the generalizability analysis, each of the facets or sources of variation in the measurement protocol were determined. Subsequently, each facet was identified as either a random or fixed effect. The day and target facets were identified and treated as random facets. That is, the conditions of these facets were identified as being a random representative sample selected from all possible observations for these facets.[12,13] For example, since the facet, days, was considered to be a random effect, then the three testing days used in the present measurement design were considered to be representative of a selected sample from all of the potential days (ie, the universe of days) from which test scores can be obtained. In contrast to the day and target facets, the 75% and 100% LOS tests were originally identified as fixed facets. The results of a preliminary analysis implementing a strategy for handling fixed facets, however, indicated that separate G studies should be performed for the two LOS tests.[13]

Variance estimates were then calculated for each source of variation in the model based on the expected mean square values and calculated mean square values from the RM ANOVA output. Thus, variance estimates were derived for subject (S), day (D), target (T), subject by day (SXD), subject by target (SXT), day by target (DXT), and the subject by day by target interaction combined with the residual random error (SXDXT, E). When performing these calculations, if negative variance components were obtained, the estimates were set to zero, and the zero value was used for any further calculations involving these variance components.[12,15,21]

A G study was then performed to assess the relative contribution of each error source as a percentage of the total measurement error, while a D Study was performed to determine the optimal measurement protocol across the universe of days and targets. G coefficients were calculated where the Days facet was generalized (ie, varied) across the three test days and the target facet was generalized across the eight test targets. The calculated G coefficients can be interpreted as reliability coefficients across the universe(s) of the various facets identified in the study.

**Table 1: Variance Components and Percentage of Variation for Movement Velocity**

| | 75% LOS | | 100% LOS | |
|---|---|---|---|---|
| Source of Variance | Variance Component | Percentage of Variation | Variance Component | Percentage of Variation |
| Subject | .574 | 18.43 | .362 | 15.01 |
| Day | .059 | 1.88 | .005 | .20 |
| Target | .524 | 16.81 | .462 | 19.18 |
| S × D | .274 | 8.81 | .083 | 3.46 |
| S × T | .352 | 11.30 | .331 | 13.75 |
| D × T | .040 | 1.29 | .000 | .000 |
| S × D × T, E | 1.291 | 41.47 | 1.166 | 48.40 |
| Total | 3.114 | 99.99 | 2.409 | 100.0 |

Abbreviations: LOS, limits of stability; S × D, subject by day; S × T, subject by target; D × T, day by target; S × D × T, E, subject by day by target combined with random error.

Additionally, the standard error of the measurement (SEM) was obtained for each of the LOS movement variables. The SEM was calculated as the square root of the absolute error variance.[8,14] The SEM provides an indication of the absolute reliability of the measure as opposed to the relative reliability provided by the G coefficient.[13,14] Specifically, the calculated SEM provides the clinician with an amount of error to be expected in a patient's performance score. The SEM values provide a confidence interval about which a subject's score is expected to vary. The lower the SEM value, relative to the mean performance score, the greater the absolute reliability of the measurement.

## RESULTS

### Day Facet

The G-study results, presented in tables 1 through 4, include the estimated variance components and percentages of variation for each of the LOS movement variables examined. As indicated in these tables, the relative variance contribution of the day facet to the total measurement error for each of the LOS movement variables was very small. Specifically, the percentage of variation in the 75% LOS test performance attributed to the day facet was less than 3% across the four variables examined. Similarly, the variance contributions of the day facet for the 100% LOS test ranged from 0% (ie, EE) to approximately 1% (ie, DC). Moreover, the summation of the variance contributions of the day facet with contributions attributed to both the SXD and DXT interactions yielded low total variances ranging from approximately 2% (ie, EE at 75% LOS test) to 12% (ie, MV at 75% LOS test).

**Table 2: Variance Components and Percentage of Variation for Maximum Excursion**

| | 75% LOS | | 100% LOS | |
|---|---|---|---|---|
| Source of Variance | Variance Component | Percentage of Variation | Variance Component | Percentage of Variation |
| Subject | 43.96 | 28.26 | 85.25 | 40.67 |
| Day | .000 | .00 | .000 | .00 |
| Target | 21.19 | 13.62 | 29.96 | 14.30 |
| S × D | 4.692 | 3.01 | 5.549 | 2.65 |
| S × T | 32.57 | 20.94 | 37.88 | 18.07 |
| D × T | .000 | .00 | .000 | .00 |
| S × D × T, E | 53.15 | 34.17 | 50.95 | 24.31 |
| Total | 155.56 | 100.00 | 209.59 | 100.0 |

Abbreviations: LOS, limits of stability; S × D, subject by day; S × T, subject by target; D × T, day by target; S × D × T, E, subject by day by target combined with random error.

**Table 3: Variance Components and Percentage of Variation for End Point Excursion**

| | 75% LOS | | 100% LOS | |
|---|---|---|---|---|
| Source of Variance | Variance Component | Percentage of Variation | Variance Component | Percentage of Variation |
| Subject | 98.05 | 22.38 | 110.86 | 29.86 |
| Day | 9.49 | 2.17 | 2.50 | .67 |
| Target | 67.70 | 15.45 | 65.23 | 17.57 |
| S × D | 8.70 | 1.99 | 3.77 | 1.02 |
| S × T | 61.69 | 14.08 | 73.50 | 19.79 |
| D × T | 1.89 | .43 | 1.92 | .52 |
| S × D × T, E | 190.56 | 43.50 | 113.52 | 30.57 |
| Total | 438.08 | 100.00 | 371.30 | 100.00 |

Abbreviations: LOS, limits of stability; S × D, subject by day; S × T, subject, by target; D × T, day by target; S × D × T, E, subject by day by target combined, with random error.

### Target Facet

In contrast to the low relative variance contributions attributed to the Day facet and the SXD and DXT interactions, the target facet and the SXT interaction accounted for larger proportions of the total variance in each of the LOS movement variables examined (tables 1 through 4). Differences in test target performance were associated with approximately 14% (ME) to 32% (DC) of the variability in the 75% LOS test scores and approximately 14% to 20% (ie, ME and DC, respectively) of the variability in the 100% LOS scores across the four dependent variables examined. Additionally, the relative variance contributions attributed to the SXT interaction ranged from approximately 11% (MV) to 21% (ME) and 14% (MV) to 27% (DC) for the 75% and 100% LOS tests, respectively.

### Residual Error Variance

A relatively large proportion of the total variance in each of the LOS movement variables was attributed to the residual error variance. The three-way interaction term (SXDXT, E) combined with random error accounted for approximately 29% to 44% (ie, DC and EE, respectively) of the variation in the 75% LOS test scores across the four LOS variables examined. Similarly, the percentages of variation associated with the SXDXT, E interaction from the 100% LOS test ranged from approximately 24% to 48% (ie, ME and MV, respectively). Thus, for both LOS tests conducted, a significant proportion of the total variability in the LOS performance was attributed to random error and/or possible sources of measurement error not included in the present G study design.

### D Study

The estimated G coefficients presented in table 5 were derived from the LOS scores on each of the three days of testing.

**Table 4: Variance Components and Percentage of Variation for Directional Control**

| | 75% LOS | | 100% LOS | |
|---|---|---|---|---|
| Source of Variance | Variance Component | Percentage of Variation | Variance Component | Percentage of Variation |
| Subject | .0019 | 12.51 | .0033 | 16.92 |
| Day | .0003 | 1.76 | .0002 | .97 |
| Target | .0050 | 32.28 | .0040 | 20.36 |
| S × D | .0004 | 2.18 | .0005 | 2.57 |
| S × T | .0032 | 20.63 | .0052 | 26.61 |
| D × T | .0002 | 1.31 | .0000 | .00 |
| S × D × T, E | .0045 | 29.33 | .0064 | 32.57 |
| Total | .0155 | 100.00 | .0196 | 100.00 |

Abbreviations: LOS, limits of stability; S × D, subject by day; S × T, subject by target; D × T, day by target; S × D × T, E, subject by day by target combined with random error.

Table 5: G Coefficient Estimates for Days = 1 Through 3 and Targets = 8

| | Movement Velocity | | Maximum Excursion | | EndPoint Excursion | | Directional Control | |
|---|---|---|---|---|---|---|---|---|
| Day | 75% LOS | 100% LOS | 75% LOS | 100% LOS | 75% LOS | 100% LOS | 75% LOS | 100% LOS |
| 1 | .54 | .57 | .74 | .84 | .71 | .80 | .60 | .63 |
| 2 | .69 | .70 | .82 | .89 | .80 | .86 | .70 | .72 |
| 3 | .75 | .75 | .85 | .91 | .84 | .88 | .73 | .75 |

Abbreviations: LOS, limits of stability.

The estimated G coefficients for the four LOS movement variables ranged from .69 (ie, MV at 75% LOS) to .89 (ie, ME at 100% LOS) for two days of testing. The addition of the third testing day yielded minimal increases in the estimated G coefficients. Collectively, the estimated G coefficients derived from the D studies indicated moderate to high reliability estimates when generalizing the LOS tests across two and three days of testing.

### Measurement Consistency

The degree of consistency in the LOS movement variables across repeated observations was determined using an $F$ test based on RM ANOVA results for both LOS tests conducted. Results from these analyses indicated no significant differences in MV ($F(2,34) = 4.56$, $p > .025$ and $F(2,23) = 2.07$, $p > .10$ for the 75% and 100% LOS tests, respectively) and ME ($F(2,25) = 1.30$, $p > .25$ and $F(2,29) = 1.02$, $p > .25$ for the 75% and 100% LOS tests, respectively) measurements across the three testing days. Similarly, for tests conducted at 100% of the theoretical stability limits, differences in EE measures across the three test days were nonsignificant ($F(2,17) = 4.50$, $p > .025$)

In contrast, significant differences in LOS performance scores across the three test days were observed for EE measures obtained for the 75% LOS test ($F(2,19) = 9.69$, $p < .005$) and for DC measures associated with both the 75% and 100% LOS tests ($F(2,20) = 6.55$, $p < .01$, and $F(2,24) = 6.56$, $p < .01$, respectively). Results of the Tukey post hoc comparisons for the EE variable indicated that EE values for day 1 were significantly lower than values obtained for test days 2 and 3. In addition, post hoc comparisons for the DC measurements indicated that DC values for days 2 and 3 were significantly lower (ie, greater COG control) than scores obtained on day 1.

Although LOS performance scores were relatively consistent across the three testing days, significant differences were observed in each of the LOS movement variables across the eight test targets. Specifically, significant target effects were evident for MV ($F(7,25) = 16.41$, $p < .001$), ME ($F(7,103) = 20.01$, $p < .001$), EE ($F(7,36) = 18.25$, $p < .01$), and DC ($F(7,100) = 21.69$, $p < .01$) for the 75% LOS test. Additionally, significant differences in MV ($F(7,45) = 28.70$, $p < .001$), ME ($F(7,96) = 22.29$, $p < .001$), EE ($F(7,56) = 19.27$, $p < .01$), and DC ($F(7,41) = 27.12$, $p < .01$) were observed for the 100% LOS test.

The results of the Tukey post hoc comparisons for the 75% and 100% LOS test performance scores indicated that the COG excursions to the forward and rear targets (target 1 and target 5, respectively) were significantly slower (ie, MV) than those for the remaining six targets. Additionally, MV scores for the left forward target (target 8) were significantly slower than those for the right forward target (target 2), right lateral target (target 3), and left rear diagonal target (target 6).

Initial COG excursions (ie, EE) were shorter within the limits of stability for both the forward and rear targets (targets 1 and 5, respectively) as compared with the remaining six targets. In contrast, EE scores for the right forward target (target 2) were significantly larger (ie, longer COG excursions through the stability region) than the values for all other targets except the left rear diagonal target (target 6).

Moreover, maximal excursions of the COG through the theoretical regions of stability (ie, ME) during the 75% LOS test were significantly smaller for both the left forward target (target 8) and rear target (target 5) when compared with the remaining six targets. Similar results in the post hoc comparisons of the ME measurements were evidenced in the 100% LOS test. The 100% LOS test results also indicated that ME values for the left forward target (target 8), rear target (target 5), and forward target (target 1) were significantly smaller than those for the remaining five targets. Results from the post hoc comparisons also indicated that the maximal excursions of the COG through the stability region for both LOS tests were largest for the right forward target (target 2), right rear diagonal target (target 4), and left rear diagonal target (target 6).

The level of COG movement control (ie, DC) to the forward target (target 1) and the rear target (target 5) was lower than the degree of control reported for the remaining six targets. In contrast, the DC values obtained for the two lateral targets (targets 3 and 7) and the right forward diagonal target (target 2) were significantly lower (ie, greater COG control) than those calculated for all other targets.

### Standard Error of Measurement

The calculated SEM values for each of the four dynamic balance measures were derived from the variance components using the full measurement protocol (ie, three test days and eight targets). The SEM values and the mean scores on each test day collapsed across the eight targets for the four LOS movement variables are presented in table 6. The calculated SEM values for each movement variable were relatively small compared with the respective mean scores.

## DISCUSSION

Collectively, the results of our investigation indicate that the LOS test, conducted at either 75% or 100% of the theoretical limits of stability, is a reliable test of dynamic balance when administered to community-dwelling, healthy older adults with no recent history of falls. Reliability estimates for MV, ME, EE, and DC using the complete LOS test (ie, eight test targets) and three test days ranged from moderately high to high.[7] Furthermore, results from the RM ANOVA indicate that, in general, the measures of dynamic balance derived from the LOS test are consistent across multiple evaluations.

Although other investigators have also reported moderately high reliability estimates for the 75% LOS test, direct comparisons with our results are not possible.[4,5] Specifically, the previous investigations reported the test-retest reliability of LOS movement variables that are no longer available in the most recent Balance Master software (version 5.0b). The original movement variables associated with the LOS test (ie, movement time, path sway, target sway, and distance error) could be problematic for both the researcher interested in establishing the test's reliability and the practitioner attempting to obtain a comprehensive assessment of dynamic balance performance across different patient populations. In previous software versions, certain movement variable scores were not provided if the subject

**Table 6: Mean Values for Days Collapsed Across the Eight Targets**

| | Movement Velocity | | Maximum Excursion | | EndPoint Excursion | | Directional Control | |
|---|---|---|---|---|---|---|---|---|
| Day | 75% LOS | 100% LOS | 75% LOS | 100% LOS | 75% LOS | 100% LOS | 75% LOS | 100% LOS |
| 1 | 2.59 | 2.67 | 103.02 | 91.37 | 86.83 | 76.90 | .304 | .295 |
| 2 | 2.76 | 2.84 | 103.75 | 92.16 | 91.57 | 78.94 | .274 | .281 |
| 3 | 3.13 | 2.85 | 104.06 | 92.44 | 93.06 | 80.47 | .271 | .265 |
| SEM | .53 | .42 | 3.24 | 3.53 | 5.50 | 4.92 | .038 | .041 |

Abbreviations: LOS, limits of stability; SEM, standard error of measurement.

did not reach a particular test target or, did not remain inside the target for a sufficient period of time (ie, 4 seconds).[18] Ceiling effects associated with the movement time variable were also evident when the individual being tested was unable to reach a test target within an 8-second time period. A maximum score of 8 seconds was recorded in these situations. These ceiling effects have the potential to bias the reliability of the test.

Unlike previous studies conducted to evaluate the reliability of certain computerized posturography tests, reliability in the present investigation was estimated using generalizability theory as opposed to an intraclass correlation analysis model. G theory is considered to be the most appropriate methodology currently available for estimating test reliability.[15] Not only does G theory extend classical reliability theory by estimating the magnitude of multiple sources of measurement error it then allows for the major sources of error to be isolated so that an optimal measurement design can be identified.[8,12,15]

### Day Facet

The results from the G studies conducted for each of the four measures of dynamic balance indicated that the day facet contributed very little to the variability in the measurement scores across the three testing days. Subsequent analyses conducted to assess the degree of consistency in the dynamic balance measures across days indicated that for three (MV, ME, and EE conducted at 100% LOS) of the four variables no statistically significant changes in the measurements were observed. In contrast to these three movement variables, subjects do not perform consistently on the DC measure until completion of the second day of testing.

Since dynamic balance measurements are both consistent and generalizable across days, as indicated by the RM ANOVA results and the low variance estimates for the day facet (ie, less than 3%), a clinician may select the recommended two LOS test sessions from a variety of test days. Although the test sessions in the present study were administered on consecutive days, previous investigations estimated the reliability of LOS tests conducted at 1-week intervals.[4,5] Similar to the present study, these investigations reported moderate to high reliability estimates of LOS performance measures.

Often in clinical practice, administration of the LOS test on multiple days may not be practical because of both time and cost constraints. Consequently, a clinician may wish to assess indices of dynamic balance on a single test day. Henderson and colleagues[5] administered multiple LOS tests within a single test session (ie, same test day). These investigators, using movement variables derived from an earlier software version, reported moderate to moderately high reliability estimates for LOS performance measures. The reported findings of Henderson and colleagues, together with the minimal variation in LOS performance measures across days in the present investigation, indicate that clinicians and researchers can expect accurate assessments of the multiple indices of dynamic balance by administering the LOS test twice on a single test day. The accuracy of the assessments on a single test day assumes, however, that the measurement characteristics of conducting multiple LOS test sessions within a single day reflect those characteristics associated with administering the LOS test across multiple days. This assumption has yet to be tested using the LOS movement variables available in the most recent Balance Master software version 5.0b.

### Target Facet

In contrast to the small variance contributions associated with the day facet, large contributions to the total measurement error were attributed to both the target facet and the interaction of subjects with targets (ie, SXT). This variability or inconsistency in the performance scores for different targets negatively affects the reliability of the dynamic balance measures as well as the generalization of the present findings from one clinical test session to another. Findings of an interactive effect for subjects by targets indicated that for both LOS tests performed, measures of dynamic balance for a particular subject varied as a function of the test target. For example, some subjects demonstrated poorer performance to the three left targets (ie, targets 6, 7, 8) as compared with other subjects.

Possible explanations forwarded to account for the variability in a subject's performance to the different targets include (1) an inability for some older subjects to produce shifts in the COG to targets positioned at 75% and 100% of the theoretical limits of stability, as well as (2) an inappropriate selection of postural strategies (eg, hip vs ankle strategy). Although many of the subjects in the present study were able to produce displacements of the COG to each of the LOS targets, for some older subjects certain target positions appear to exceed their actual limits of stability. These findings are consistent with recent investigations that have reported age-related declines in the region of stability.[4,22] Thus, the variability in the subjects by targets error component may be an indication of such age-related declines in the actual LOS for some of the older adult subjects tested in the present study.

### Unexplained Variance

An additional error source providing large contributions to the variability in the total measurement was the highest order interaction (SXDXT, E) combined with the random error component. This error component contains not only the unexplained random variance but also the error variance attributed to facets that have not been identified in the study. In the present investigation, age may have been a potential facet contributing to the variability in the dynamic balance measurements. Age effects have been previously identified by Hageman and colleagues[6] as a factor influencing an individual's ability to control the speed and accuracy of COG movements within the region of stability.

Although standardization of the test instructions minimizes the variability associated with the subjects' understanding of the task to be performed, misinterpretations of the visual biofeedback may still occur. Specifically, some of our subjects may have experienced difficulties understanding the information provided by the COG cursor, despite standardized test instructions

and the opportunity to adequately familiarize themselves with the visual biofeedback. Consequently, the potential misunderstanding of the verbal instructions and/or the visual biofeedback may have contributed to the variability in the dynamic balance measures attributed to the random error component.

### Standard Error of the Measurement

In contrast to relative differences in the measurements provided by generalizability theory analysis, the SEM reflects the absolute amount of measurement error expected.[7,10] Thus, for the clinician, the SEM may be considered a more practical measure of reliability. Small SEM values relative to the mean scores, as reported in the present investigation, indicate that a limited range of possible performance scores should be expected on subsequent evaluations of a patient.

### Clinical Implications

For clinicians who wish to perform tests of balance or evaluate patient progress during the course of a balance intervention program, knowledge of the expected sources of measurement error and the consistency of balance measures across multiple testing sessions can provide meaningful information. For example, the low variance estimates for the day facet reported in the present study indicate that dynamic balance measures associated with the LOS test are generalizable and consistent across days. Consequently, clinicians may confidently administer multiple evaluations of the LOS test on a variety of possible days. Relative error variances associated with targets and the subjects by targets interaction further indicate the importance of performing the complete LOS test (ie, testing all eight targets). The reliability (ie, generalizability) increased as the number of test targets evaluated increased. Clinicians should therefore avoid the tendency to conduct an abbreviated version of the test in order to minimize evaluation time. Conducting the complete test also provides a more comprehensive evaluation of the patient's region of stability and will better assist the clinician in identifying specific movement limitations.

One final implication emerging from our findings is the importance of providing clear test instructions and sufficient practice time for patients to better understand the relationship between the movement of the on-screen COG cursor and the actual movements of the body's COG. It is recommended that patients be allowed to explore movements of the cursor on a blank screen prior to actually administering the test. The type of postural strategy adopted may be more consistent as a result of such practice and more representative of the patient's actual dynamic balance abilities.

### Study Limitations and Conclusions

Results of this investigation indicate that the 75% and 100% LOS tests were reliable tests of dynamic balance when administered to healthy older adults with no recent history of falls. Although we recognize that the LOS test is not administered to healthy adult populations in a clinical setting, we considered it necessary first to establish the reliability of the LOS test, featuring a new array of performance variables, with a nonpatient population. The reliability of these tests are currently being investigated with a variety of patient populations (eg, asymptomatic older adult fallers, hemiparetic patients, etc).

This study also showed that multiple evaluations of dynamic balance measures were generally consistent across two and three days of testing with relatively low variance contributions attributed to the day facet. Thus, to obtain measures of dynamic balance that are both reliable (ie, generalizable) and consistent, a minimum of two LOS test sessions is recommended during which the complete LOS test is performed. It remains to be determined, however, if these recommendations hold true for different patient populations evaluated on this test of dynamic balance.

**Acknowledgments:** The authors thank the participants for their involvement in this study. Additionally, we thank Dr. Terry Wood (Oregon State University) for his suggestions and contributions regarding data analyses.

### References

1. Shephard NT, Telian SA, Smith-Wheelock M, Raj A. Vestibular and balance rehabilitation therapy. Ann Otol Rhinol Laryngol 1993; 102:198-205.
2. Chester JB. Whiplash, postural control, and the inner ear. Spine 1991;16:716-20.
3. Ford-Smith CD, Wyman JF, Elswick Jr. RK, Fernandez T, Newton RA. Test-retest reliability of the sensory organization test in noninstitutionalized older adults. Arch Phys Med Rehabil 1995;76:77-81.
4. Liston RAL, Brouwer BJ. Reliability and validity of measures obtained from stroke patients using the balance master. Arch Phys Med Rehabil 1996;77:425-30.
5. Henderson NE, Overby AS, Panzer VP. Internal consistency and stability of balance measures among different age groups [abstract]. In: Proceedings of the 12th International Congress of the World Confederation for Physical Therapy: 1995 June 25-30; Washington, DC. Alexandria (VA): The American Physical Therapy Association Inc; 1995. p. 751.
6. Hageman PA, Leibowitz M, Blanke D. Age and gender effects on postural control measures. Arch Phys Med Rehabil 1995;76:961-5.
7. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. Norwalk (CT): Appleton and Lange; 1993.
8. Morrow Jr. JR. Generalizability theory. In: Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics; 1989. p. 73-96.
9. Mitchell SK. Interobserver agreement, reliability and generalizability of data collected in observational studies. Pychol Bull 1979;86: 376-90.
10. Thomas JR, Nelson JK. Introduction to research in health, physical education, recreation and dance. Champaign (IL): Human Kinetics; 1985. p. 253-68.
11. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: John Wiley & Sons Inc.; 1972.
12. Brennan RL. Elements of generalizability theory. Iowa City: American College Testing Program; 1983.
13. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. Am Psychol 1989;44:922-32.
14. Roebroeck ME, Hariaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993;73:386-401.
15. Shavelson RJ, Webb NM. Generalizability theory: a primer. Newbury Park (CA): Sage Publications Inc.; 1991.
16. Lahey MA, Downey RG, Saal FE. Intraclass correlations: there's more than meets the eye. Psychol Bull 1983;93:586-95.
17. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
18. NeuroCom International, Inc. Balance Master operators manual. Clackamas (OR): NeuroCom International, Inc.; 1993.
19. Crick JE, Brennan RL. GENOVA: a general purpose analysis of variance system [computer program]. Dorchester (MA): University of Massachusetts at Boston, Computer Facilities; 1984.
20. Myers JL, Wells AD. Research design and statistical analysis. New York: Harper-Collins; 1991.
21. Cardinet J, Tourneur Y, Allel L. Extension of generalizability theory and its applications in educational measurement. J Educ Meas 1981; 13:183-204.
22. Blaszczyk JW, Lowe DL, Hansen PD. Ranges of postural stability and their changes in the elderly. Gait Posture 1994;2:11-7.

### Supplier

a. NeuroCom International, Inc., 9570 SE Lawnfield Road, Clackamas, OR 97015-9611.