AWARD NUMBER: W81XWH-17-1-0547

TITLE: Can a novel beam-walking test improve fall risk assessment in Service members, Veterans, and civilians who use lower-limb prostheses?

PRINCIPAL INVESTIGATOR: Andrew Sawers, CPO, PhD

CONTRACTING ORGANIZATION: University of Illinois at Chicago
CHICAGO IL 60612

REPORT DATE: October 2019

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.
Title: Can a novel beam-walking test improve fall risk assessment in Service members, Veterans, and civilians who use lower-limb prostheses?

Authors: Andrew Sawers

Abstract:
The purpose of the proposed research is to improve the assessment of fall risk among unilateral lower limb prosthesis users. The scope of the proposed research includes establishing the validity and reproducibility (test-retest and inter-rater reliability) of a new clinical balance test, the Narrowing-Beam Walking Test (NBWT). This is to be accomplished by administering the NBWT and four contemporary performance-based balance tests to 60 lower limb prosthesis users, who are then followed prospectively for 6 months to identify fall events. To date we have completed recruitment and enrollment of study participants at both study sites (UIC and UW), and are well into data analysis. Dissemination efforts have resulted in the publication of four manuscripts, five invited talks or symposia, and six posters or podium presentations. Four more manuscripts are currently in review or preparation. Our Narrowing Beam Walking Test has also been adopted and implemented by a prosthetic manufacturer (Ohio Willow Wood), and several academic/clinical sites (e.g., Minneapolis VA, Center for the Intrepid, Seattle VA, and Delft University). Our research was also featured in the CDMRP FY19 Annual Report.

Subject Terms:
amputee; balance; falls; recruitment; enrollment

Distribution Statement: Approved for Public Release; Distribution Unlimited
Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2. Keywords</td>
<td>4</td>
</tr>
<tr>
<td>3. Accomplishments</td>
<td>4-9</td>
</tr>
<tr>
<td>4. Impact</td>
<td>9</td>
</tr>
<tr>
<td>5. Changes/Problems</td>
<td>9</td>
</tr>
<tr>
<td>7. Participants &amp; Other Collaborating Organizations</td>
<td>11</td>
</tr>
<tr>
<td>8. Special Reporting Requirements</td>
<td>12</td>
</tr>
<tr>
<td>9. Appendices</td>
<td>12</td>
</tr>
</tbody>
</table>
INTRODUCTION

Existing clinical balance tests are too easy. They do not pose sufficient challenge to expose subtle, but critically important differences in balance that underlie fall risk in Service members, Veterans, and civilians lower limb prosthesis users. Consequently they are unable to discriminate between lower limb prosthesis users with and without a history of falls, or predict who is at risk for a fall. As a result there are no effective or accepted screening tests for diagnosing fall risk in lower limb prosthesis users. To address this gap and improve the assessment of fall risk in lower limb prosthesis users, the purpose of this project is to assess the validity and reliability of a novel Narrowing Beam-Walking Test (NBWT) that provides progressively increasing challenge to balance control. The scope of the proposed research includes establishing the criterion validity (Aim 1), reproducibility (test-retest and inter-rater reliability) (Aim 2), and predictive validity (Aim 3) of a new clinical balance test, the Narrowing-Beam Walking Test (NBWT). This is to be accomplished by administering the NBWT and four contemporary performance-based balance tests to 60 lower limb prosthesis users. These individuals will then be followed prospectively for 6 months to report fall events. This project addresses a notable gap in the DoD/VA lower limb amputation rehabilitation clinical practice guidelines, a valid and reliable test to predict fall risk in lower limb prosthesis users.

KEYWORDS

Amputee; falls; fall risk; balance; prosthesis; prospective; validity; reliability; rehabilitation

ACCOMPLISHMENTS

Major goals of the project: The major goals (milestones) of the project as outlined in the SOW were:

Goal 1: Obtain and maintain IRB approvals from UIC, UW, and ORP/HRPO (Target date: Y1Q1, Y1Q4, Y2Q4 – 100% Completed).

Goal 2: Recruitment, consent, and data collection materials/databases/equipment/staff ready for data collection (Y1Q1- 100% Completed).

Goal 3: 60 participants enrolled (Y2Q2 – 100% completed).

Goal 4: Data collection completed (Y2Q3 – 100% completed).

Goal 5: Data entered, processed, and analyzed to address study hypotheses (Y2Q4 – 80% completed).

Goal 6: Abstracts presented at a minimum of 2 scientific conferences, manuscripts prepared and submitted for publication, delivery of training material for clinical implementation of beam-walking test, presentation of executive summary to senior VA/DoD clinical personnel (Y2Q4 – 75% completed).
Accomplishments under these goals: For each of the goals/milestones outlined in the SOW, we have made significant and timely progress beyond our Y1 Annual Report.

Goal 1: Both study sites (UIC and UW) have completed their annual IRB continuing review (2019). Documents for the ORP/HRPO annual continuing review were submitted in Oct 2019.

Goal 2: Completed in previous reporting period (Y1).

Goal 3: We have met all of our quarterly enrollment targets at both study sites. We have recruited n=60 participants between the UIC site and the UW site. Because of our success recruiting lower limb amputees for the project we will not have to enhance or expand our recruitment efforts at either study site.

Goal 4: Planned data collection has been completed at both study sites. Namely, clinical balance test data and prospective 6-month fall data have been collected on all 60 participants. Our success in recruitment and data collection has allowed us to extend the prospective follow up period from 6 to 12 months. This will allow us to: i) evaluate “how long” fall risk predictions are good for by comparing predictions over 6 and 12 months, and ii) assess fall recall among people with lower limb amputation, topics we have found to be poorly addressed within the existing literature.

Goal 5: All data has been entered, via double entry to minimize data entry errors, processed, and is currently being analyzed.

Goal 6: We continue to maintain a productive publication record under this award. To date we have published four peer-reviewed articles, 3 during the current reporting period. An additional four manuscripts are either under review or in preparation. All of these manuscripts present results that are relevant to the current project. The first article published during this reporting period, “Conventional administration and scoring procedures suppress the diagnostic accuracy of a performance-based test designed to assess balance ability in lower limb prosthesis users”, was published in Prosthetics and Orthotics International. The main result from this paper was that if practice effects on the Narrowing Beam Walking Test were not accommodated through administration and scoring procedures (i.e., taking the mean of trials 3 to 5 rather than the best of the first two trials), the diagnostic accuracy of the Narrowing Beam Walking Test for fall risk was significantly lower (i.e., the area under the Receiver Operating Characteristic curve dropped from .81 to .67). The take home message from this paper was therefore that the effectiveness of the Narrowing Beam Walking Test at discriminating between fallers and non-fallers was not due to the challenge it presents to balance control, as was originally hypothesized, but rather having accommodated practice effects via administration and scoring procedures. These results suggest that practice effects in other performance-based clinical balance tests may similarly limit their diagnostic accuracy.

The second manuscript, “Using clinical balance tests to assess fall risk in established lower limb prosthesis users: cutoff scores and related validity indices”, was published in PM & R. In this paper we introduced cut-off scores for a suite of performance-based clinical balance tests that facilitate the discrimination between fallers and non-fallers (Table 1). These cut-off scores were accompanied by a number of validity indices, including likelihood ratios, which clinicians can use to assess the
The third manuscript, "Inter-rater and test-retest reliability of performance-based clinical tests administered to unilateral lower limb prosthesis users," is in press at Physical Therapy. In this paper we derived reliability indices that clinicians can use to select balance tests that are suitable for probability that an individual patient will fall in the next 12 months (Table 1). Such validity indices were previously absent from the literature, limiting clinicians' ability to assign an individual patient probability of falling. The assignment of a probability is an improvement over previously available metrics that provide a binary outcome, faller or non-faller.

### Table 1. Validity indices for performance-based clinical balance tests among unilateral lower limb prosthesis users for ≥2 falls (multiple falls)

<table>
<thead>
<tr>
<th>Test</th>
<th>CI, AUC (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR- (95% CI)</th>
<th>Sensitivity (95%)</th>
<th>Specificity (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT</td>
<td>2.77 (12.63)</td>
<td>0.1 (0.1-3.1)</td>
<td>0.1 (0.3-7.1)</td>
<td>73% (63%-86%)</td>
<td>7% (5%-13%)</td>
</tr>
<tr>
<td>Model 1: NBWT; NBW; FG; MVC</td>
<td>0.92 (1.9-9.6)</td>
<td>0.1 (0.05-0.5)</td>
<td>0.05 (0.02-0.1)</td>
<td>90% (85%-95%)</td>
<td>9% (7%-11%)</td>
</tr>
<tr>
<td>Model 2: NBWT; NBW; FG; MVC</td>
<td>0.96 (1.9-9.6)</td>
<td>0.1 (0.05-0.5)</td>
<td>0.05 (0.02-0.1)</td>
<td>90% (85%-95%)</td>
<td>9% (7%-11%)</td>
</tr>
<tr>
<td>NBWT</td>
<td>2.77 (12.63)</td>
<td>0.1 (0.1-3.1)</td>
<td>0.1 (0.3-7.1)</td>
<td>73% (63%-86%)</td>
<td>7% (5%-13%)</td>
</tr>
<tr>
<td>Model 1: NBWT; NBW; FG; MVC</td>
<td>0.92 (1.9-9.6)</td>
<td>0.1 (0.05-0.5)</td>
<td>0.05 (0.02-0.1)</td>
<td>90% (85%-95%)</td>
<td>9% (7%-11%)</td>
</tr>
<tr>
<td>Model 2: NBWT; NBW; FG; MVC</td>
<td>0.96 (1.9-9.6)</td>
<td>0.1 (0.05-0.5)</td>
<td>0.05 (0.02-0.1)</td>
<td>90% (85%-95%)</td>
<td>9% (7%-11%)</td>
</tr>
</tbody>
</table>

**NBWT:** Narrowing Beam Walking Test; **FG:** Forward Step Test; **MVC:** Maximum Voluntary Contraction; **AUC:** Area Under the Curve; **CI:** Confidence Interval; **LR+:** Positive Likelihood Ratio; **LR-:** Negative Likelihood Ratio; **Model 1:** NBWT; NBW; FG; MVC; **Model 2:** NBWT; NBW; FG; MVC.
making group-level comparisons or individual-level clinical decisions (i.e., test-retest intra-class correlation coefficients), ii) interpret the precision of individual test scores on performance-based clinical balance tests (i.e., standard error of measurement), and iii) evaluate change over time in balance performance among individual lower limb prosthesis users (i.e., minimal detectable change) (Table 2). These reliability indices were previously absent from the literature, limiting the ability of clinicians to select and interpret performance-based clinical balance tests among lower limb prosthesis users.

Table 2. Test-retest reliability of performance-based clinical balance tests in people with lower-limb loss

<table>
<thead>
<tr>
<th>Test</th>
<th>Session 1 Score Mean (SD)</th>
<th>Session 2 Score Mean (SD)</th>
<th>Score Change Mean (SD)</th>
<th>ICC (2,1) (95% CI)</th>
<th>SEM</th>
<th>MDC90 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (score, /1.00)</td>
<td>.42 (.21)</td>
<td>.44 (.24)</td>
<td>.02 (.10)</td>
<td>.90 (.83-.94)</td>
<td>.07</td>
<td>.16 (.10-.20)</td>
</tr>
<tr>
<td>TUG (time, s)</td>
<td>10.4 (2.76)</td>
<td>9.91 (2.75)</td>
<td>.50 (1.30)*</td>
<td>.88 (.78-.93)</td>
<td>.96</td>
<td>2.2 (1.5-2.9)</td>
</tr>
<tr>
<td>FSST (time, s)</td>
<td>10.4 (5.27)</td>
<td>10.3 (4.83)</td>
<td>.14 (1.32)</td>
<td>.97 (.94-.98)</td>
<td>.87</td>
<td>2.0 (.79-3.5)</td>
</tr>
<tr>
<td>10MWT (time, s)</td>
<td>8.74 (1.96)</td>
<td>8.58 (2.00)</td>
<td>.16 (.54)</td>
<td>.96 (.93-.98)</td>
<td>.39</td>
<td>.91 (1.40-1.4)</td>
</tr>
</tbody>
</table>

* Test score significantly different between test sessions (2-sided paired t-test) p=0.004. NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement, MDC90: Minimal Detectable Change

We have also submitted a manuscript that details the magnitude, time course, and impact of practice effects on performance-based clinical balance tests when administered to lower limb prosthesis users. We found that practice effects (i.e., significant changes in performance between consecutive trials), were observed among all administered tests (i.e., Timed Up and Go (TUG), Four Square Step Test (FSST), and the 10 meter Walk Test (10mWT)). Practice effects on the FSST, TUG, and 10mWT were observed in 76% (16/21), 68% (13/19), and 45% (9/20) of participants respectively. Performance stabilized (i.e., practice effects dissipated) on each test for all participants within 9 trials (Figure 1).

Figure 1. Example trial-by-trial data from two LLP users, one with practice effects (dark), and one without practice effects (white).

Later stable performance on the FSST and TUG was significantly better (i.e., times were lower) than conventionally derived scores that do not accommodate practice effects (i.e., best of first two trials) (Figure 2). The difference in conventional scores between test session 1 and test session 2 (7 days apart) frequently exceed the minimal detectable change of each performance-based balance test. In contrast, differences in later stable performance between test session 1 and test session 2 did not (Figure 3). Specifically, despite no change in health or prosthetic prescription, conventional scoring procedures resulted in between session differences exceeding the minimal detectable change of the FSST and TUG.
in 24% and 11% of participants respectively. These results suggest that statistically and clinically significant differences in balance ability appear to be obstructed when conventional administration and scoring procedures are used with the FSST and TUG. This may contribute to their limited ability to detect change and accurately discriminate between fallers and non-fallers. Additional research is required to develop suitable modifications to accommodate or resolve practice effects in these tests.

Three other manuscripts are currently in preparation. These include: i) the predictive validity of the Narrowing Beam Walking Test, and threshold cut-off scores to prospectively predict fallers (Aim 3); ii) the accuracy of fall recall among people with lower limb amputation (new question), and iii) “how long” are fall risk predictions “good for” in people with lower limb amputation (i.e., how often should fall risk be re-assessed) (new question).

Opportunities for training and professional development: Nothing to report.
Dissemination of results to communities of interest: Results were disseminated to a national audience of prosthetists at the 2019 American Academy of Prosthetists and Orthotists Annual Meeting and Scientific Conference, and the 2019 Orthotics and Prosthetics Innovative Technologies Conference, to local prosthetists and lower limb prosthesis users at the 2019 Scheck and Siress Education Fair, national researchers at the 2019 International Society of Posture and Gait Research (ISPGR), and the 2019 American Congress of Rehabilitation Medicine (ACRM), as well as military and civilian rehabilitation experts at the 2019 MHSRS conference.

Plan to accomplish goals during over next reporting period: During our no cost extension, we intend to complete all remaining facets of the study goals. This includes: i) analyzing all remaining data pertaining to Aim 3 and additional questions we have developed; ii) conducting hypothesis testing for Aim 3 and additional questions we have developed; iii) prepare and submit manuscripts that: a) present the predictive validity of the Narrowing Beam Walking Test, and threshold cut-off scores to prospectively predict fallers (Aim 3), b) evaluate the accuracy of fall recall among people with lower limb amputation (new question), and c) assess “how long” fall risk predictions are “good for” in people with lower limb amputation (new question). We will also develop training materials (e.g., written manual, instructional video) for the Narrowing Beam Walking Test that we will make available online, and author and present an executive summary of the project results to the VA Amputation System of Care, and VA/DOD Extremity Trauma and Amputation Center of Excellence.

IMPACT: The performance-based balance test that we have designed, developed, and tested, the Narrowing Beam Walking Test, is now being used in a number of clinical, industry, and academic centers. This includes but is not limited to: Ohio Willow Wood, the Minneapolis VA Health Care System (Adaptive Design and Engineering Program), Center for the Intrepid, VA Puget Sound Health Care System (Center for Limb Loss and Mobility), and Delft University.

CHANGES/PROBLEMS
None to report.

PRODUCTS
Journal publications (in this reporting period)


Conference presentations (in this reporting period)
Sawers A, Hafner, BJ. Ossur Instructional Course. Increasing awareness of knowledge around falls in lower limb prosthesis users. Reykjavik, Iceland, 2019 (*invited speaker*).


Sawers, A. Northwestern University’s Neuromechanics of Rehabilitation for Lower Limb Loss Symposium. Falls in lower limb prosthesis users: Refocusing through an epidemiological lens. Chicago, IL, 2018 (*invited speaker*).


Kim J, Major MJ, Hafner BJ, Sawers A. Characterizing fall circumstances as reported by lower limb prosthesis users: A secondary analysis. American Congress of Rehabilitation Medicine 96th Annual Conference Chicago, IL 2019 (*Poster*).
**PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS**

Individuals who worked on the project

<table>
<thead>
<tr>
<th>Name</th>
<th>Project Role</th>
<th>Researcher Identifier</th>
<th>Nearest person month worked</th>
<th>Contribution to Project</th>
<th>Funding Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew Sawers, PhD</td>
<td>Principal Investigator (UIC)</td>
<td></td>
<td>1</td>
<td>Dr. Sawers has been responsible for overseeing all aspects of the project (IRB, managing recruitment, enrollment, and data collection, manuscript preparation)</td>
<td>N/A</td>
</tr>
<tr>
<td>Brian Hafner, PhD</td>
<td>Co-Investigator (UW)</td>
<td></td>
<td>1</td>
<td>Dr. Hafner has been responsible for overseeing aspects of the project at the UW study site (IRB, managing recruitment, enrollment, and data collection, manuscript preparation)</td>
<td>N/A</td>
</tr>
<tr>
<td>Janis Kim, MPT</td>
<td>Graduate Student (UIC)</td>
<td></td>
<td>1</td>
<td>Ms. Kim has performed work in the area of screening of participants, enrolling participants, consent, data collection, data entry, data analysis, and dissemination at the University of Illinois at Chicago study site. She has also managed the prospective fall follow up phone calls at the UIC study site.</td>
<td>N/A</td>
</tr>
<tr>
<td>Geoff Balkman</td>
<td>Graduate Student (UW)</td>
<td></td>
<td>&lt;1</td>
<td>Mr. Balkman has been responsible for data collection and data entry at the University of Washington study site.</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Change in the active other support for the PD/PI(s) or senior/key personnel
Nothing to report

Other organizations involved as partners
Nothing to report

SPECIAL REPORTING REQUIREMENTS
N/A

APPENDICES
Appendix 1: Quad Chart
Appendix 2: Manuscripts
Can a novel beam-walking test improve fall risk assessment in Service Members, Veterans, and civilians who use lower limb prostheses?

OP160027
W81XWH-16-OPORP-PORA (Funding Level 1)
PI: Andrew Sawers, PhD, CPO
Org: The University of Illinois at Chicago
Award Amount: $300,000

Study Aim(s)
• Assess and compare the criterion validity of beam walking in lower limb prosthesis (LLP) users to existing balance tests.
• Assess and compare the reliability of beam walking in LLP users to existing clinical balance tests.
• Assess and compare the predictive validity of beam walking in LLP users to existing clinical balance tests.

Approach
We will conduct cross-sectional (Aim 1) and longitudinal (Aims 2 & 3) studies to establish the validity and reliability of a novel beam walking-test to assess fall risk in LLP users. We will compare the ability of the beam-walking test with existing clinical balance tests to retrospectively discriminate between LLP users with and without a history of falls (Aim 1), and prospectively classify fallers and non-fallers over a 6 month follow-up period (Aim 3). The reliability of the beam-walking test will be tested between raters and days (Aim 2).

Timeline and Cost

<table>
<thead>
<tr>
<th>Activities</th>
<th>CY 17</th>
<th>CY 18</th>
<th>CY 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human subjects approval &amp; train sites</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Participant recruitment</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Conduct data collection procedure</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Analyze data and disseminate results</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

Estimated Budget ($K) $300 $75 $150 $75

Updated: (01/07/2010)

Accomplishment: Conventional scoring procedures resulted in between session differences (i.e., test-retest) that exceeded the minimal detectable change (MDC90) for the Four Square Step Test (A), Timed Up and Go (B), and 10-meter Walk Test (C). When scored using conventional procedures, the difference between test and retest performance exceeded or was not significantly different from the MDC of the FSST, TUG, and 10mWT in 5/21 (24%), 2/19 (11%), and 0/20 (0%) of participants respectively (A-C (i)). In contrast, scoring the FSST and TUG based on later stable performance reduced the number of participants whose between session differences exceeded or were no different than the MDC to 0/21 (0%) on the FSST, 0/19 (0%) on the TUG, and 0/19 (0%) on the 10mWT (A-C (ii)). Similar changes were not observed with the 10mWT.

Goals/Milestones

CY17 Goal – Study preparation and participant recruitment
☑ Obtain human subjects approval
☑ Equip and train study staff at UIC & UW

CY18 Goals – Ongoing recruitment, data collection, and analysis
☑ Recruit additional 40 participants
☑ Conduct data collection and analysis procedures
☑ Disseminate initial results at national conference

CY19 Goal – Analysis and dissemination
☑ Recruit and collect data from final 10 participants
☑ Analyze final data set
☐ Disseminate final study results

Comments/Challenges/Issues/Concerns
• None to report

Budget Expenditure to Date
$207,500 (including F&A)
Conventional administration and scoring procedures suppress the diagnostic accuracy of a performance-based test designed to assess balance ability in lower limb prosthesis users

Andrew Sawers1 and Brian J Hafner2

Abstract

Background: Practice effects have been observed among performance-based clinical tests administered to prosthesis-users. Their impact on test applications remains unknown.

Objective: To determine whether scoring a clinical balance test using conventional procedures that do not accommodate practice effects reduces its diagnostic accuracy relative to scoring it using recommended procedures that do accommodate practice effects.

Study Design: Cross-sectional study.

Methods: Narrowing Beam Walking Test data from 40 prosthesis users was scored using recommended methods (i.e. average of trials 3–5) and conventional methods applied to other tests (i.e. mean or best of trials 1–3). Area under the receiver operating characteristic curve for each method was compared to 0.50, to determine if it was better than chance at identifying prosthesis-users with a history of falls, and to 0.80, to determine if it surpassed a threshold recommended for diagnostic accuracy.

Results: Receiver operating characteristic curve area decreased when the Narrowing Beam Walking Test was scored using conventional rather than recommended procedures. Furthermore, when scored using conventional procedures, the NBWT no longer discriminated between prosthesis-users with and without a history of falls with a probability greater than chance, or exceeded recommended diagnostic thresholds.

Conclusion: Scoring the Narrowing Beam Walking Test using conventional procedures that do not accommodate practice effects decreased its diagnostic accuracy among prosthesis-users relative to recommended procedures. Conventional scoring procedures may limit the effectiveness of performance-based tests used to screen for fall risk in prosthesis-users because they do not mitigate practice effects. The influence of practice effects on other tests, and test applications (e.g. clinical evaluation and prediction), is warranted.

Clinical relevance

Scoring a clinical balance test using conventional procedures that do not mitigate practice effects reduced its diagnostic accuracy. Changing administration and scoring procedures to accommodate practice effects should be considered to improve the diagnostic accuracy of other performance-based balance tests.

Keywords

Amputee, accidental falls, psychometrics, rehabilitation

Date received: 25 January 2019; accepted: 12 April 2019
Background

Over 50% of lower limb prosthesis (LLP) users report falling at least once a year, placing them at a high risk for adverse health outcomes that include injury, reduced mobility, and diminished quality of life. Falls are therefore a considerable problem that negatively affect the lives of a substantial portion of LLP users. A major barrier to reducing falls among LLP users has been effective screening of those at risk. Existing performance-based clinical balance tests do not generally pose sufficient challenge to identify differences in balance control that underlie fall risk, reducing their ability to discriminate between LLP users with and without a history of falls.

A standardized performance-based clinical balance test termed the Narrowing Beam Walking Test (NBWT) was developed to address this limitation among LLP users. The NBWT was designed to safely challenge and detect differences in balance across a wide range of abilities by restricting the base-of-support and/or reducing the width of the support surface. The NBWT requires tested individuals walk along four low beams (2.5 cm height, 1.8 m long), each narrower than the last (20 cm, 10 cm, 4.0 cm, and 2.0 cm). Validation testing among LLP users demonstrated that the NBWT was the only one of several clinical balance tests to discriminate between LLP users with and without a history of falls with a probability greater than chance, and above a recommended threshold for diagnostic accuracy. The accuracy with which the NBWT discriminated between LLP users with and without a history of falls was attributed to the increased challenge beam walking imposes on balance control.

Beyond increasing the challenge to balance control, development of the NBWT included the study practice effects associated with beam-walking performance by LLP users. Practice effects are systematic changes in performance that occur between consecutive trials or repeated assessments with the same test in the absence of any intervention. Practice effects need to be accommodated (e.g. by allowing tested individuals to perform several practice trials) to avoid taking measurements during the time when individuals are improving on the test. Although practice effects have been observed in performance-based tests among LLP users and other patient populations, few clinical test developers explicitly state whether their recommended administration and scoring procedures address possible practice effects, and none have examined the role practice effects may have on the discriminative, evaluative, or predictive applications of performance-based clinical tests.

During development of the NBWT, investigators used data from 10 consecutive trials to evaluate practice effects in LLP users, and create an administration and scoring procedure (i.e. the mean of trials 3 through 5) that accurately estimated LLP users’ consistent (i.e. typical) performance. The developers, however, did not compare the diagnostic accuracy of NBWT scores obtained with their recommended administration and scoring procedure to scores obtained using conventional methods for scoring clinical balance tests (e.g. mean or best of trials 1–3). It remains unknown therefore whether the superior diagnostic accuracy of the NBWT among LLP users is due to the challenge the NBWT imposes on balance control, as originally hypothesized, or use of an administration and scoring procedure that mitigates practice effects.

The primary objective of this study was to determine whether scoring the NBWT using methods generally recommended for other clinical balance tests (i.e. mean or best of trials 1–3) would affect its ability to discriminate LLP fallers and non-fallers. It was hypothesized that scoring the NBWT using conventional scoring methods would adversely affect its diagnostic accuracy (i.e. reduce the area under the receiver operating characteristic (ROC) curve). To test this hypothesis, the diagnostic accuracy of NBWT scores derived using three different administration and scoring procedures—mean of trials 1–3, mean of trials 1–3, and best of trials 1–3 were compared. Significant differences in diagnostic accuracy would indicate that practice effects limit the effectiveness of the NBWT, and that additional research examining whether practice effects influence the diagnostic accuracy of similar performance-based balance tests may be required. A secondary objective was to determine whether factors previously associated with practice effects in performance-based clinical tests (e.g. baseline performance, age, level of motor deficit) could explain practice effects observed among LLP users performing the NBWT.

Methods

Study design

A secondary analysis of cross-sectional data from a sample of 40 LLP users collected in a previous study was conducted to address the study objectives. Original study protocols were reviewed and approved by institutional review boards at the University of Illinois at Chicago. All individuals provided written informed consent prior to participation.

Participants

Inclusion criteria for the prior study were as follows: 18 years of age or older; unilateral transtibial or transfemoral amputation due to trauma, dysvascular complications, tumor, or infection; one year or more of experience using a prosthesis, regular use of a prosthesis to ambulate; and ability to walk 10 feet over level terrain with the prosthesis and without an assistive device. Exclusion criteria included amputation of a second limb, complications on the contralateral leg (e.g. knee replacement, ulcers), severe pulmonary disease, an advanced neurologic disorder, or congenital limb absence or difference. People with congenital limb absence or difference were excluded because their motor
control, and potential balance, may be fundamentally different than someone who losses a limb later in life.

Measurements

Participant demographic and characterization measures. Sociodemographic information (i.e. height, weight, age, and sex) was collected via self-report. Medicare Functional Classification Level (MFCL) (i.e. K-level was determined by a certified prosthetist via interview and physical evaluation, while amputation-related information (i.e. amputation level, amputation etiology, and time since amputation) was obtained via interview with a study investigator. To identify participants with a history of recent falls, each was asked, “In the past year have you experienced any falls including a slip or trip where you inadvertently lost your balance and landed on the ground or lower level?”28–30 Participants reporting one or more falls in the past 12 months were classified as “fallers.”8,31–33

Assessment of practice effects in narrowing beam walking. Participants’ performance across 10 narrowing beam-walking trials was examined in a prior study in order to develop administration and scoring procedures for the NBWT.11 Performance on each trial was quantified by the distance walked prior to the participant stepping off the beam or uncrossing their arms, whichever came first. Practice effects were identified as statistically significant changes in the slope of each participant’s cumulative trial-by-trial performance.11,34 Participants’ narrowing beam-walking performance was considered stable and consistent once the cumulative trial-by-trial record reached a terminal slope (i.e. the participant exhibited no further changes in performance across trials).11 Participants were classified into one of two groups; LLP users who did and did not experience significant changes in performance over the 10 beam walking trials (i.e. participants who did or did not exhibit practice effects). The distance walked along the NBWT on the first trial was selected as a measure of baseline performance.

An administration and scoring procedure was then developed to mitigate the observed narrowing beam-walking practice effects,11 and ensure that the NBWT assessed balance ability based on an accurate estimate of tested individuals’ stable and consistent performance. Specifically, the mean of trials 3–5 was found to more accurately estimate stable consistent narrowing beam-walking performance than methods used by most contemporary clinical balance tests (i.e. best or mean of trials 1–3).11 The former method was therefore recommended for administering and scoring the NBWT (11).

Statistical analysis

Distributions of NBWT scores derived using each scoring method were evaluated for normality using the Shapiro-Wilk test. A one-way repeated measures analysis of variance (ANOVA) was used to compare NBWT scores across the three scoring methods (i.e. mean of trials 3–5, mean of trials 1–3, best of trials 1–3). Diagnostic accuracy of each method was evaluated using the area under the respective receiver operating characteristic (ROC) curve.35 Areas were first compared to a null value of 0.50 to assess whether the procedure was better than chance at identifying participants with a history of falls. Areas were then compared to a threshold of 0.80, a recommended threshold for clinically acceptable diagnostic accuracy.36 The significance level was adjusted with a Bonferroni correction (α < .0167) to account for multiple comparisons. To test whether factors previously linked to practice effects (i.e. age, baseline performance, motor deficits)15,27 differed between LLP users who did and did not exhibit practice effects on the NBWT, continuous variables (e.g. age, baseline performance) were compared between groups using 2-sided t-tests, while nominal variables (e.g. amputation level, fall history) were compared using Fisher’s exact tests. Statistical analyses were performed using SPSS v.24 (Chicago, IL).

Results

Narrowing beam-walking data from 40 LLP users (Table 1) who participated in the original study8 were analyzed. Practice effects, lasting between one and seven trials, were observed among 17 participants. NBWT scores (mean ± 95% confidence interval (CI)) for the mean of trials 3–5 (0.40 ± 0.13), the best of trials 1–3 (0.44 ± 0.15), and the mean of trials 1–3 (0.36 ± 0.14) were all normally distributed (mean 3–5: W = 0.964, p = 0.24; best 1–3: W = 0.973, p = 0.45; mean 1–3: W = 0.959, p = 0.21). Mauchly’s test of sphericity indicated that the assumption of sphericity was violated, χ²(2) = 6.508,
Table 2. Participant characteristics potentially contributing to practice effects.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>NBWT baseline first trial</th>
<th>MFCL*</th>
<th>Amputation level*</th>
<th>Amputation etiology*</th>
<th>Fall history*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice effects (n = 17)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>K1&amp;2</td>
<td>K3&amp;4</td>
<td>TT</td>
</tr>
<tr>
<td>49.35 (12.52)</td>
<td>0.405 (0.232)</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>No practice effects (n = 23)</td>
<td>48.26 (16.17)</td>
<td>0.290 (0.209)</td>
<td>10</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>p-value</td>
<td>0.818</td>
<td>0.109</td>
<td>1.00</td>
<td>0.187</td>
<td>0.107</td>
</tr>
</tbody>
</table>

MFCL: Medicare Functional Classification Level (K-level); TT: transtibial; TF: transfemoral; DV: dysvascular; NDV: non-dysvascular.
*Statistical significance was assessed with a Fisher’s Exact Test.

Results show that the diagnostic accuracy of the NBWT was 36% lower when scored using methods that do not accommodate practice effects. When the NBWT was scored using conventional methods, it was no longer able to discriminate between fallers and non-fallers with a probability greater than chance (i.e. the best of trials 1–3), or exceed an established threshold for diagnostic accuracy (i.e. mean and best of trials 1–3). These results indicate that accommodating practice effects, rather than the challenge posed by NBWT, may be responsible for its superior diagnostic accuracy. Furthermore, these results suggest that the influence of practice effects on the diagnostic ability of other performance-based clinical tests should be investigated. This may have important implications for the use of performance-based clinical tests in clinical trials, as well as day-to-day clinical practice.

**Practice effects influence the diagnostic accuracy of the NBWT.** Several studies have documented practice effects in performance-based clinical tests.15–20 We are not aware, however, of any research that has assessed the influence of practice effects on the discriminative, evaluative, or predictive ability of performance-based clinical tests. The results of this study therefore contribute several key results to the body of knowledge. First, the diagnostic accuracy of the NBWT dropped to suboptimal levels reported for other clinical balance tests applied among LLP users when it was scored in a manner used by other contemporary balance tests (i.e. best of trials 1 and 2).22–25 Second, the NBWT’s ability to identify people with a history of falls dropped below a recommended threshold of accuracy when it was scored in this manner. Third, the NBWT was unable to identify LLP users with a history of falls better than chance when it was scored using the best of trials 1–3. In contrast, when the NBWT is scored using its recommended methods that mitigate practice effects (i.e. mean of trials 3–5), its diagnostic accuracy was both significantly better than chance and surpassed the recommended threshold for diagnostic accuracy.

The current results, while focused on one performance-based balance test, represent an important, albeit initial, step in understanding how practice effects can limit the effectiveness of clinical tests. Presence of practice effects in performance-based tests indicates that an individual’s performance is not consistent, but is instead changing over repeated administrations (i.e. due to familiarity with the test or practice). The consequence of practices effects, whether in clinical practice or research, is that an individual’s initial baseline measurements may not represent their true ability. Correspondingly, changes in measured performance (e.g. after an intervention or over time) may not be accurate. Thus, tests that exhibit practice effects may give erroneous information about whether or how the individual has changed. Examples of tests that may exhibit practice effects in LLP users have begun to appear in Halsne et al.16

Whereas previous studies have advocated increasing the challenge to balance control as a way to improve the diagnostic accuracy of clinical tests,9,37 these results suggest that mitigating practice effects may also facilitate discrimination.
between LLP users with and without a history of falls. Had increasing the challenge to balance control alone been sufficient, the NBWT scored using the best or mean of trials 1–3 would have had an ROC curve area greater than 0.80. Rather, it appears that it is necessary to mitigate practice effects, in this case via administration and scoring procedures, in order to achieve an acceptable level of diagnostic accuracy. The ability of the NBWT to discriminate between LLP users with and without a history of falls with greater accuracy than existing clinical balance tests8 may therefore be attributed not only to increasing the challenge to balance control,9,10 but also to mitigating practice effects through administration of additional trials.11 Interestingly, practice effects associated with narrowing beam walking appear to affect either an overestimation (i.e. the best of trials 1–3), or underestimation (i.e. the mean of trials 1–3) of individuals’ consistent NBWT performance among LLP users. Controlling or reducing the variance attributable to practice effects appears to be a critically important consideration when using performance-based clinical balance tests like the NBWT to evaluate fall risk among LLP users. Whether similar results would be observed with other performance-based clinical balance tests remains to be determined.

The cause of practice effects may dictate how they ought to be addressed and potentially interpreted. Practice effects in performance-based clinical tests have been associated with individuals who are younger,15,27 exhibit fewer motor deficits,15,27 demonstrate better baseline performance,26 and receive greater verbal encouragement.15 In this study, practice effects on the NBWT could not be explained by differences in age, baseline performance, mobility, or level of motor deficit (i.e. amputation level, etiology) (Table 2). Thus, it would not be feasible to address practice effects by adjusting NBWT scores based on these factors. Rather, developing appropriate administration and scoring procedures would appear to be the most suitable approach to resolving practice effects.

While practice effects in a few performance-based clinical tests have been related to a number of factors described above,15,26,27 the reasons for practice effects on a given test remain largely unknown. Considering potential explanations for practice effects in performance-based clinical tests may have important implications for how they are identified, interpreted, and best addressed. For example, if practice effects are related to issues of motivation, effort, or movement strategy selection, they may be best addressed by modifying test instructions. If specific motor tasks were found to be susceptible to practice effects, then they could be removed from existing clinical tests and avoided when developing new tests. Each of these solutions, however, assumes an understanding of practice effects and their causes that has yet to be established.

While the focus of this study was on resolving practice effects, practice effects themselves may provide important and useful information about the integrity and adaptability of the systems mediating an individual’s performance. For example, LLP users who exhibited practice effects in the current study walked further on their first NBWT trial, reported fewer falls, and had a greater NBWT score (i.e. mean of trials 3–5) (Table 2). While these results did not reach statistical significance, they do suggest the possibility that LLP users who can adapt their gait, measured here through practice effects, may be safer ambulators. Additional research is required to examine this hypothesis in a larger sample. Depending on the setting therefore, practice effects may be conceptualized either as a source of unwanted variance needing to be removed, or as another factor to be considered in clinical decisions.38

Study limitations. The current study only assessed the influence of practice effects on one performance-based clinical test (i.e. the NBWT), one application (i.e. discrimination), one construct (i.e. balance), and one clinical population (i.e. LLP users). Whether similar results would be observed with other balance tests (e.g. timed performance tests, clinical rating scales), applications (i.e. prediction and evaluation), constructs (i.e. mobility and endurance), and/or clinical populations (e.g. incomplete spinal cord injury, stroke, older adults) remains to be determined. Practice effects observed in other performance-based clinical tests,15,17,18 including those applied to LLP users,16,19,20 suggest that the current results extend beyond this study. The sample size used in the present analysis (n = 40), while consistent with similar studies,23,39 should be increased in future work to confirm the current results and add greater certainty to their interpretation. Finally, only a limited set of variables was studied in an effort to explain why practice effects may be occurring. Studied variables were limited to participant demographics and characterization measures collected in the original study.8 The collection of physiological measures of mood, effort, attention, cognition, and movement strategies in future studies may yield further insights into the basis of practice effects among performance-based clinical tests.

Conclusion

The results of the present study suggest that the information obtained from performance-based clinical tests may be limited by how they are administered and scored. Rescoring the NBWT using methods often used with other clinical tests significantly decreased its diagnostic accuracy. The ability of the NBWT to discriminate between LLP users with and without a history of falls with greater accuracy than existing clinical balance tests can therefore be attributed not only to having imposed greater challenge to balance control, but also to mitigating practice effects. In light of their prominent use in clinical trials and clinical practice, results of this study suggest that consideration should be given to whether the discriminative, predictive, or evaluative applications of other performance-based clinical tests are influenced by practice effects.
Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the National Institutes of Health (NIH) under award number K12HD073945, by the Orthotics and Prosthetics Education and Research Foundation, Inc. (OPERF) under grant number OPERF- SGA-2016-1, and by the Department of Defense (DoD) under award number W81XWH-17-1-0547. This content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, OPERF, or DoD.

Supplemental material
Supplemental material for this article is available online.

ORCID iDs
Andrew Sawers https://orcid.org/0000-0002-3493-304X
Brian J Hafner https://orcid.org/0000-0001-6175-1869

References


Original Research

Using Clinical Balance Tests to Assess Fall Risk among Established Unilateral Lower Limb Prosthesis Users: Cutoff Scores and Associated Validity Indices

Andrew Sawers, PhD, CPO, Brian J. Hafner, PhD

Abstract

Background: Clinicians are routinely required to make decisions about fall risk among lower limb prosthesis (LLP) users. These decisions can be guided by standardized clinical balance tests but require population- and test-specific cutoff scores and validity indices to categorize individuals as probable fallers or nonfallers on the basis of test performance. Despite the importance of cutoff scores and validity indices to clinical interpretation of clinical balance test scores, they are rarely reported for LLP users. In their absence, clinicians cannot use results from clinical balance tests to assess the likelihood of a fall by any one patient.

Objective: Derive cutoff scores, and associated validity indices, for clinical balance tests administered to established unilateral LLP users.

Design: Cross-sectional study.

Setting: Outpatient clinic and research laboratory.

Participants: Established ambulatory unilateral transtibial and transfemoral prosthesis users (n = 40).

Intervention: Not applicable.

Main Outcome Measure(s): Optimal cutoff scores and related validity indices (ie, area under the curve, sensitivity, specificity, likelihood ratios) were computed for five balance tests, the activities-specific balance confidence scale (ABC), timed up and go (TUG), four square step test (FSST), Berg balance scale (BBS), and narrowing-beam walking test (NBWT).

Results: Cutoff scores were identified for the NBWT (≤0.43/1.0), TUG (≥8.17 seconds), FSST (≥8.49 seconds), BBS (≤50.5/56), and ABC (≤80.2/100). Validity indices (ie, area under the curve, sensitivity, specificity, and likelihood ratios) for the NBWT, TUG, and FSST had greater diagnostic accuracy and provided more information about the probability of a fall than those for the BBS or ABC.

Conclusion: Performance above or below identified cutoff scores for the NBWT, FSST, and TUG provides information about potentially important shifts in the probability of falling among established unilateral LLP users. These results can serve as initial benchmarks to reduce uncertainty surrounding fall risk assessment in established unilateral LLP users but require future prospective evaluation.

Level of Evidence: III

Introduction

Clinicians are routinely required to make important decisions about whether a patient may be at risk for an adverse condition or event. Diagnostic decisions such as these are ideally made using psychometrically sound clinical tests with which the probability of an event or condition occurring or existing can be estimated. Establishing population- and test-specific cutoff scores and validity indices is key to applying clinical tests in this manner. Cutoff scores are required to dichotomize continuous scales and create benchmarks that categorize individuals as with or without the condition or event on the basis of a “positive” or “negative” test result. Validity indices (eg, sensitivity, specificity, and likelihood ratios) serve to provide information on the probability of the event or condition. In the absence of cutoff scores and associated validity indices, clinical tests can be administered to evaluate change over time or differences between individuals but cannot be used to assess...
Fall Risk Assessment in Lower Limb Prosthesis Users

Falls remain a frequent event that negatively affects the lives of a substantial portion of lower limb prosthesis (LLP) users.\(^3\)\(^,\)\(^4\) Over 50% of LLP users report falling at least once a year, with up to 39% reporting multiple falls a year.\(^4\)\(^-\)\(^8\) Falls among LLP users frequently lead to adverse health outcomes including injury, financial costs, reduced mobility, and diminished quality of life.\(^1\)\(^4\) A major barrier to reducing falls among LLP users has been effective screening of those at risk.\(^1\)\(^5\)\(^,\)\(^6\) Central to this barrier is a scarcity of LLP user-specific cutoff scores and validity indices for contemporary clinical balance instruments such as the Berg balance scale (BBS),\(^1\)\(^7\) timed up and go (TUG),\(^1\)\(^8\) four square step test (FSST),\(^1\)\(^9\) and activities-specific balance confidence (ABC) scale.\(^2\)\(^0\)

Two different faller classifications, either \(\geq 1\) fall\(^3\)\(^,\)\(^1\)\(^5\)\(^,\)\(^2\)\(^8\),\(^2\)\(^9\) or \(\geq 2\) falls\(^1\)\(^,\)\(^5\)\(^,\)\(^2\)\(^1\)\(^)\(^)\(^,\)\(^2\)\(^7\) have been used to assess the diagnostic accuracy and validity indices of performance-based clinical balance tests. Validity indices for one performance-based balance test, the BBS, increased when a classification of \(\geq 2\) falls was used in a prior study.\(^8\) Adopting a \(\geq 2\) falls classification inherently creates a “faller” group that is likely to possess worse balance ability than if it also included individuals with 1 fall (ie, \(\geq 1\) fall classification). As a result, it may be easier to identify fallers versus nonfallers and possibly make it more difficult to identify people with moderate balance impairments who may be at risk for falls. Despite both faller classifications being used in research and clinical settings, no study of LLP users has systematically evaluated the influence of faller classification on validity indices associated with performance-based clinical balance tests commonly administered to established unilateral LLP users.

The objective of this study was therefore to establish cutoff scores, and associated validity indices, for clinical balance tests administered to established unilateral LLP users. In prior studies it was found that single-task tests like the FSST, TUG, or narrowing beam walking test (NBWT) demonstrated better discriminant validity than multi-item tests like the BBS, ABC, or the Locomotor Capabilities Index.\(^1\)\(^6\)\(^,\)\(^2\)\(^1\)\(^)\(^)\(^)\(^,\)\(^6\)\(^)\(^,\)\(^2\)\(^7\) Based on this prior research, therefore, it was hypothesized that the validity indices (eg, likelihood ratios) associated with cutoff scores for the FSST, TUG, and the NBWT would exceed those of the BBS or ABC scale. Confirmation of this hypothesis would offer further evidence that clinical tests like the NBWT, FSST, and TUG can be used to assess fall risk among established unilateral transtibial and transfemoral prosthesis users. Further, cutoff scores and related validity indices would provide clinicians with the information needed to appropriately interpret clinical balance test scores to assess fall risk in their patients who use LLPs. Owing to the variety of ways researchers and clinicians can classify individuals as fallers or nonfallers,\(^3\)\(^,\)\(^1\)\(^5\)\(^,\)\(^1\)\(^6\)\(^,\)\(^2\)\(^1\)\(^)\(^)\(^,\)\(^2\)\(^7\) a secondary objective of this study was to determine whether cutoff scores and associated validity indices differed across two common fall classifications (ie, \(\geq 1\) fall vs \(\geq 2\) falls over the past 12 months). It was hypothesized that classifying fallers as \(\geq 2\) falls would increase validity indices.\(^8\)

Methods

Study Design

A cross-sectional study was performed from July 2016 to May 2017. The STROBE (Strengthening The Reporting of Observational studies in Epidemiology) Statement guideline\(^3\)\(^0\) was followed during the collection and reporting of study data. All data were stored and managed using a REDCap database hosted at the University of Illinois at Chicago.\(^3\)\(^1\) Study protocols were reviewed and approved by a University of Illinois at Chicago institutional review board. Study participants provided written informed consent prior to participation.

Participants

LLP users were recruited from local prosthetic clinics. Inclusion criteria included unilateral transtibial or transfemoral amputation due to traumatic, dysvascular, or oncologic causes; 18 years of age or older; one or more years of using a prosthesis (ie, established users); ability to ambulate at least 10 ft without an upper extremity assistive device (eg, cane); and use of a comfortable prosthesis (assessed with the Socket Comfort Scale).\(^3\)\(^2\) Participants with complications to their contralateral leg (eg, joint replacement, arthritis, or wounds), amputation of a second limb, or a neurological or cardiovascular condition that limited the ability to complete the study protocol were excluded.

Procedures

Participants completed demographic and prosthetic-related characterization measures, as well as a retrospective falls survey. Wearing their preferred prosthesis-footwear combination, participants were administered five clinical balance tests. Five-minute rest periods were enforced between each balance test. Cutoff scores and validity indices were derived using recommended methods.\(^1\)\(^,\)\(^3\)\(^3\)\(^-\)\(^3\)\(^5\)
Measurements

Participant Demographic and Characterization Measures

Age, height, weight, and sex were collected from study participants via self-report. Medicare functional classification level (MFCL) (ie, K-level) was determined by a certified prosthetist via interview and physical evaluation, and amputation-related information (ie, level, etiology, and time since) was obtained via interview with a study investigator. Perceived mobility was assessed with the Prosthetic Limb Users Survey of Mobility (PLUS-M). To ascertain the number of falls experienced by each participant over the past 12 months, participants were asked, “In the past year have you had any falls including a slip or a trip in which you inadvertently lost your balance and landed on the ground or lower level?” Participants who reported falling in the past year were then asked to recall the number of falls in the past 12 months. To determine whether different faller classifications influence cutoff scores and related validity indices, data were analyzed using two faller classifications: ≥1 fall and ≥2 falls over the past 12 months.

Clinical Balance Tests

Five clinical balance instruments, the ABC Scale, the TUG, the FSST, the BBS, and the NBWT were administered and scored according to the developers’ instructions (Appendix 1). Each balance instrument has demonstrated acceptable levels of validity and/or reliability (ie, intraclass correlation coefficient range .70-.99) among LLP users.

Statistical Analysis

Clinical balance test scores were compared across participant subgroups defined by their reported fall history (ie, 0 falls, 1 fall, or ≥2 falls), and within commonly adopted fall classifications (ie, 0 falls vs ≥1 fall, or 0-1 falls vs ≥2 falls) using parametric (ie, analysis of variance) or nonparametric equivalents (Kruskal–Wallis Test).

Receiver operating characteristic (ROC) curves were used to identify cutoff scores and related validity indices (ie, area under the curve, sensitivity, specificity, and likelihood ratios) for each clinical balance test. ROC curves were obtained by plotting the sensitivity of a test against 1-specificity. The area under the curve (AUC) represents the probability of correctly identifying a faller from a randomly selected pair of lower limb prosthesis users, one being a faller and the other a nonfaller. The larger the AUC, the greater the test’s general discriminative ability.

The AUC was selected as a summary measure of diagnostic accuracy. The AUC can assume any value between 0 and 1, with a value of 0.5 representing chance, and values greater than or equal to 0.8 recommended as the limit of clinical acceptability. Areas for each test were therefore compared to a threshold of 0.8 to determine the clinical acceptability of a test. Although the AUC provides insight into the overall discriminant ability of a test based on group data, it does not provide clinicians with actionable information when assessing an individual patient. Cutoff scores and validity indices such as likelihood ratios are required to make individual-level decision based on test scores.

Optimal cutoff scores were chosen for each balance test by selecting the point on a test’s ROC curve closest to the (0,1) point. This point was identified by choosing the minimal value of the function (1-sensitivity)² + (1-specificity)², a method that maximizes sensitivity and specificity. The test score corresponding to this minimal value was therefore selected as the cutoff score that best discriminates between unilateral LLP users with and without a history of falls. Values for sensitivity, specificity, and specificity, and thereby, do not indicate the probability of the event or condition occurring, information that is likely to be of greatest value to clinicians. Predictive values alternatively offer probabilities of an event occurring but are dependent on the prevalence of the event in the study sample and thus rarely generalize beyond the study. Likelihood ratios overcome these aforementioned limitations by quantifying how much the obtained test score increases or decreases the probability of an event occurring, independent of its prevalence in the sample (ie, they generalize beyond the study). Likelihood ratios are therefore considered more efficient and clinically useful than sensitivity and specificity values or positive and negative predictive values. Unlike the AUC, likelihood ratios provide clinicians with actionable information for individual patients (ie, the probability a patient will or will not fall based on a positive or negative test outcome). Likelihood ratios were therefore selected as the primary validity index in this study. The likelihood ratio for a positive test (LR+) was computed as sensitivity/1-specificity, whereas the likelihood ratio for a negative test (LR-) was computed as 1-specificity/sensitivity. Likelihood ratios greater than 5 or less than 0.2 indicate moderate changes in the probability of an event (eg, fall), while likelihood ratios between 5 and 2, or 0.5 and 0.2 result in small, but potentially important shifts in probability. Likelihood ratios greater than 2 or less than 0.2 were considered the minimum for considering a clinical test able to contribute to a fall risk assessment.

A multivariate logistic regression model was developed for each faller classification (ie, ≥1 fall = Model 1, ≥2 falls = Model 2) to determine whether a combination of clinical balance test scores and other fall-related variables could enhance validity indices beyond the score from any single test. Univariate logistic regression was first performed with each potential predictor variable.
(eg, FSST time, age, sex, etiology, level of amputation, mobility level) as the independent variable, and faller status as the dependent variable. Independent variables found to have an odds ratio of ≥2 or <0.5, and a P value of <.05 were retained for further consideration. To avoid multicollinearity, any of the independent variables retained from the univariate logistic regression that had a strong correlation (ie, Spearman rho ≥ .75) and a lower odds ratio than the other independent variable with which it was correlated were excluded from the multivariate logistic regression model. A multivariate logistic regression was then performed with all remaining independent variables included. Validity indices described previously were computed from the regression output. Cut-off scores and validity indices were computed for each of the faller classifications (ie, ≥1 fall and ≥2 falls). 95% confidence intervals (CI) were computed to determine the precision of each validity index. All statistical analyses were conducted using SPSS v.25 (SPSS, Inc., Chicago, IL).

Results

The sample included 40 established unilateral LLP users (height: 173 ± 9.10 cm; weight: 78.5 ± 14.1 kg; socket comfort score [SCS]: 7.8 ± 1.4) (Table 1). All participants were able to perform and complete all tests. Scores for each of the studied tests, stratified by participant fall history and fall classification are presented in Tables 2 and 3, respectively. Pairwise comparisons revealed that with the exception of the ABC, participants with a history of 2 or more falls had significantly worse test scores than those with a history of either one fall or no falls (P < .027 to .004) (Table 2). There were no statistically significant differences in test scores between participants who reported 1 and 0 falls. When a fall classification scheme of 2 or more versus 0-1 falls in the past 12 months was implemented, pairwise comparisons revealed that again with the exception of the ABC, participants reporting 2 or more falls had worse test scores than those reporting 0-1 falls (Table 3). With the exception of the NBWT, these differences in test scores between groups were not observed when a fall classification scheme of 1 or more versus 0 falls was implemented (Table 3).

Cutoff scores and validity indices (estimate, 95% CI) for each clinical balance test are presented in Table 4. When fallers were defined as individuals reporting ≥1 falls in the 12 months prior to assessment, the NBWT had the greatest area under the ROC curve (AUC) (0.81, 0.62-0.91), as well as the largest specificity (76%, 54%-96%) and positive likelihood ratio (3.0, 1.5-6.9). The TUG had the greatest sensitivity (83%, 68%-98%) and smallest negative likelihood ratio (0.24, 0.13-0.56). The logistic regression model developed for the ≥1 fall classification (ie, Model 1) included the NBWT, FSST, PLUS-M, and amputation level as predictor variables. Many, but not all, validity indices improved slightly compared to individual clinical balance test scores alone (Table 4). Other candidate predictor variables (eg, TUG, ABC, MFCL, amputation etiology, age and sex) failed to meet model inclusion criteria (ie, odds ratio ≥2 or less than 0.5, and P < .05), during univariate logistic regression.

When fallers were defined as individuals reporting ≥2 falls in the 12 months prior to assessment validity indices (ie, sensitivity, specificity, and likelihood ratios) associated with each of the clinical balance tests generally improved (Table 4). The NBWT again had the largest AUC (0.89, 0.78-1.0). The FSST, however, exhibited the largest sensitivity (94%, 83%-100%) and smallest negative likelihood ratio (0.08, 0.012-0.54), whereas the TUG had the greatest specificity (83%, 67%-98%) and positive likelihood ratio (4.7, 1.9-11.9). The logistic regression model developed for the ≥2 falls classification (ie, Model 2) included several predictor variables consistent with those included in Model 1 (ie, NBWT, PLUS-M), and others that were unique to Model 2 (ie, TUG, MFCL, and amputation etiology). Model 2 did not improve validity indices as markedly as Model 1 did (Table 2). Model 2 was not among the top three results for several of the validity indices (ie, sensitivity and negative likelihood ratio).

Discussion

The objective of this study was to establish cutoff scores and associated validity indices for several clinical balance tests administered to established unilateral LLP users. Results supported the hypothesis that validity indices, including likelihood ratios, of the NBWT, FSST, and TUG exceeded those for the BBS or ABC scale. Performance above or below identified cutoff scores for the NBWT, FSST, and TUG appear to provide information about small but potentially important shifts in the

Table 1
Participant demographics stratified by faller status

<table>
<thead>
<tr>
<th>Fall Group</th>
<th>Age (y) Mean (SD)</th>
<th>Sex</th>
<th>Years since Amputation Mean (SD)</th>
<th>Amputation Level</th>
<th>Amputation Etiology</th>
<th>PLUS-M Mean (SD)</th>
<th>MFCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No falls</td>
<td>44.9 (14.4)</td>
<td>M (11) F (5)</td>
<td>16.4 (12.8)</td>
<td>TT (14) TF (2)</td>
<td>Dysvascular (1) Nondysvascular (15)</td>
<td>58.5 (9.05)</td>
<td>K1 (0) K2 (4) K3 (10) K4 (2)</td>
</tr>
<tr>
<td>1 fall</td>
<td>41.4 (16.3)</td>
<td>M (2) F (5)</td>
<td>10.8 (4.8)</td>
<td>TT (3) TF (4)</td>
<td>Dysvascular (0) Nondysvascular (7)</td>
<td>56.2 (6.35)</td>
<td>K1 (0) K2 (1) K3 (2) K4 (4)</td>
</tr>
<tr>
<td>≥2 falls</td>
<td>55.4 (11.7)</td>
<td>M (8) F (9)</td>
<td>13.8 (14.8)</td>
<td>TT (8) TF (9)</td>
<td>Dysvascular (6) Nondysvascular (11)</td>
<td>49.2 (6.71)</td>
<td>K1 (2) K2 (10) K3 (5) K4 (0)</td>
</tr>
</tbody>
</table>

PLUS-M = Prosthetic Limb User’s Survey of Mobility; TF = Transfemoral; TT = Transtibial; M = Male; F = Female; Other = trauma, cancer, infection; MFCL = Medicare Functional Classification Level.
Tables

**Table 2**

Clinical balance test scores stratified by fall history

<table>
<thead>
<tr>
<th>Fall Status</th>
<th>NBWT (/1.0) Mean (SD)</th>
<th>TUG (s) Median (IQR)</th>
<th>FSST (s) Median (IQR)</th>
<th>BBS (/56) Median (IQR)</th>
<th>ABC (/100) Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 falls (n = 16)</td>
<td>0.51 (.03) - .27-.76</td>
<td>7.86 (2.83) 5.99-13.0</td>
<td>7.56 (4.87) 5.26-12.8</td>
<td>52.0 (8.50) 40.0-55.0</td>
<td>90.3 (18.9) 63.1-100</td>
</tr>
<tr>
<td>1 fall (n = 7)</td>
<td>0.53 (.06) -.25-.74</td>
<td>7.45 (3.04) 5.06-9.18</td>
<td>6.57 (3.08) 4.14-11.8</td>
<td>52.0 (6.00) 50.0-56.0</td>
<td>90.6 (14.1) 53.1-99.2</td>
</tr>
<tr>
<td>≥2 falls (n = 17)</td>
<td>0.25 (.06) -.04-.63</td>
<td>10.0 (5.89) 8.30-19.0</td>
<td>13.8 (6.76) 6.27-39.3</td>
<td>45.0 (12.5) 16.0-55.0</td>
<td>74.26 (9.96)</td>
</tr>
<tr>
<td>Total (n = 40)</td>
<td>0.41 (0.19) -.04-.76</td>
<td>8.75 (3.97) 5.06-19.0</td>
<td>8.83 (5.94) 4.14-39.3</td>
<td>50.0 (8.50) 16.0-56.0</td>
<td>85.0 (23.7) 40.9-100</td>
</tr>
</tbody>
</table>

NBWT = Narrowing Beam Walking Test; TUG = Timed Up and Go; FSST = Four Square Step Test; BBS = Berg Balance Scale; ABC = Activities-specific Balance Confidence Scale; SD = Standard Deviation; IQR = Interquartile Range.

**Table 3**

Clinical balance test scores stratified by two faller classifications

<table>
<thead>
<tr>
<th>Fall Status</th>
<th>NBWT (/1.0) Mean (SD)</th>
<th>TUG (s) Median (IQR)</th>
<th>FSST (s) Median (IQR)</th>
<th>BBS (/56) Median (IQR)</th>
<th>ABC (/100) Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 fall (any falls)</td>
<td>0.51 (.13) -.27-.76</td>
<td>7.86 (2.83) 5.99-13.0</td>
<td>7.56 (4.87) 5.26-12.8</td>
<td>52.0 (8.50) 40.0-55.0</td>
<td>90.3 (18.9) 63.1-100</td>
</tr>
<tr>
<td>≥2 falls (multiple falls)</td>
<td>0.25 (.06) -.04-.63</td>
<td>10.0 (5.89) 8.30-19.0</td>
<td>13.8 (6.76) 6.27-39.3</td>
<td>45.0 (12.5) 16.0-55.0</td>
<td>74.26 (9.96)</td>
</tr>
</tbody>
</table>

NBWT = Narrowing Beam Walking Test; TUG = Timed Up and Go; FSST = Four Square Step Test; BBS = Berg Balance Scale; ABC = Activities-specific Balance Confidence Scale; SD = Standard Deviation; IQR = Interquartile Range.

**Table 4**

Validity indices for performance-based clinical balance tests among unilateral lower limb prosthesis users for two faller classifications

<table>
<thead>
<tr>
<th>Test</th>
<th>AUC (95% CI)</th>
<th>Cutoff Score</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>LR- (95% CI)</th>
<th>LR+ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (/1.00)</td>
<td>0.81 (.62-.91)</td>
<td>≤0.43</td>
<td>73% (53%-90%)</td>
<td>76% (54%-96%)</td>
<td>3.0 (1.5-6.9)</td>
<td>0.36 (19-77)</td>
</tr>
<tr>
<td>TUG (s)</td>
<td>0.71 (.54-.88)</td>
<td>≥0.17</td>
<td>83% (68%-98%)</td>
<td>68% (46%-92%)</td>
<td>2.6 (1.3-5.6)</td>
<td>0.24 (13-56)</td>
</tr>
<tr>
<td>FSST (s)</td>
<td>0.70 (.53-.86)</td>
<td>≥0.49</td>
<td>74% (58%-92%)</td>
<td>68% (46%-92%)</td>
<td>2.4 (1.1-5.2)</td>
<td>0.36 (17-78)</td>
</tr>
<tr>
<td>BBS (/56)</td>
<td>0.66 (.47-.83)</td>
<td>≥0.50</td>
<td>67% (48%-86%)</td>
<td>62% (39%-86%)</td>
<td>1.8 (0.89-3.6)</td>
<td>0.53 (27-11)</td>
</tr>
<tr>
<td>ABC (/100)</td>
<td>0.65 (.47-.82)</td>
<td>≤0.80</td>
<td>50% (30%-70%)</td>
<td>74% (54%-96%)</td>
<td>1.9 (.78-5.1)</td>
<td>0.67 (41-11)</td>
</tr>
<tr>
<td>Model 1</td>
<td>0.84 (.71-.96)</td>
<td>N/A</td>
<td>80% (64%-96%)</td>
<td>73% (51%-96%)</td>
<td>3.0 (.13-.71)</td>
<td>0.27 (.12-.63)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>AUC (95% CI)</th>
<th>Cutoff Score</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>LR- (95% CI)</th>
<th>LR+ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (/1.00)</td>
<td>0.89 (.78-.91)</td>
<td>≤0.43</td>
<td>88% (73%-100%)</td>
<td>79% (65%-92%)</td>
<td>4.2 (1.7-6.9)</td>
<td>0.16 (.042-60)</td>
</tr>
<tr>
<td>TUG (s)</td>
<td>0.88 (.78-.98)</td>
<td>≥0.25</td>
<td>82% (64%-100%)</td>
<td>83% (67%-98%)</td>
<td>4.7 (1.9-11.9)</td>
<td>0.21 (.075-61)</td>
</tr>
<tr>
<td>FSST (s)</td>
<td>0.88 (.76-.99)</td>
<td>≥0.71</td>
<td>94% (83%-100%)</td>
<td>74% (60%-92%)</td>
<td>3.6 (1.8-7.3)</td>
<td>0.08 (0.12-54)</td>
</tr>
<tr>
<td>BBS (/56)</td>
<td>0.81 (.66-.92)</td>
<td>≥0.50</td>
<td>88% (73%-100%)</td>
<td>70% (51%-88%)</td>
<td>2.9 (1.5-5.5)</td>
<td>0.17 (0.45-64)</td>
</tr>
<tr>
<td>ABC (/100)</td>
<td>0.71 (.54-.87)</td>
<td>≤0.80</td>
<td>65% (42%-87%)</td>
<td>78% (61%-95%)</td>
<td>3.0 (1.3-7.0)</td>
<td>0.45 (23-89)</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.90 (.79-.10)</td>
<td>N/A</td>
<td>76% (56%-97%)</td>
<td>83% (67%-98%)</td>
<td>4.4 (1.7-11.3)</td>
<td>0.28 (.12-68)</td>
</tr>
</tbody>
</table>

ABC = Activities-specific Balance Confidence Scale; AUC = Area Under the Curve; BBS = Berg Balance Scale; CI = Confidence Interval; FSST = Four Square Step test; LR- = negative likelihood ratio; LR+ = positive likelihood ratio; NBWT = Narrowing Beam Walking Test; TUG = Timed Up and Go.

**Model 1**: NBWT; FSST; PLUS-M; amputation level.

**Model 2**: NBWT; TUG; PLUS-M; K-level; amputation etiology.

The probability of falling among unilateral LLP users. The distribution of demographic (eg, age), amputation (eg, level, etiology, and time since), and activity characteristics (eg, PLUS-M, MFCL) of participants in the present study were generally comparable to those reported in large national studies of people with lower limb amputation (n = 210-1568), although skewed slightly toward a higher percentage of individuals with non-dysvascular and transtibial amputation. The overall similarity, however, with these samples suggests that the...
current results can be generalized to the broader population of established unilateral LLP users. Although these results can serve as initial benchmarks to reduce uncertainty surrounding the assessment of fall risk in established unilateral LLP users, clinicians and researchers should also consider the reliability and ease of use of these instruments when deciding whether to adopt them in their clinical or academic practices. Additionally, prospective evaluation of the cutoff scores and their validity indices will be required in future studies.

This study provides the first set of cutoff scores and validity indices for performance-based clinical balance tests based on data collected from established unilateral transtibial and transfemoral prosthesis users. The current results build on the work of Dite et al,21 who established cutoff scores and validity indices for several clinical balance tests (eg, FSST, TUG) using data from a sample of short-term (ie, 6 months postdischarge) transtibial prosthesis users. Notable differences were observed between the cutoff scores in the present study and those reported by Dite et al21. Dite et al21 used a faller classification of ≥2 falls and reported cutoffs of 19 and 24 seconds for the TUG and FSST, respectively. Cutoff times of 9.25 and 8.71 seconds were identified for the TUG and FSST in the present study. Times were even lower (ie, 8.17 and 8.49 seconds, respectively) if the “≥1 fall” classification was used. The discrepancies in cutoff times noted between the two studies may be attributed to the study samples. The present study included both transtibial and transfemoral prosthesis users, whereas Dite et al21 included only transtibial prosthesis users. However, the higher prevalence6 and risk of falls59 among transfemoral prosthesis users would be expected to increase cutoff times for the FSST and TUG in the present study relative to the prior study, not decrease them. Also, previous studies have not reported statistically significant differences in TUG16,22 or FSST16 times between transtibial and transfemoral prosthesis users. Thus, inclusion of transtibial and transfemoral prosthesis users in the present study cannot explain the observed differences in cutoff scores. The observed differences in cutoff times are more likely to be attributed to differences in time since amputation. Dite et al21 studied participants less than a year after amputation (ie, 6.4 ± 1.5 months after discharge from the rehabilitation unit), whereas participants in the current study averaged 14.3 years since amputation. This difference implies that distinct cutoff scores may be required to evaluate fall risk at different times after amputation and that balance ability and fall risk may change markedly after 6 months of prosthesis use. Longitudinal changes of balance and fall risk may be an important area of future research. The results of these two studies should therefore be considered complimentary rather than conflicting, serving two temporally distinct groups; short-term versus established LLP users.

Existing clinical balance tests appear to provide important information about the probability of falls in established unilateral LLP users. Although none of the likelihood ratios for the clinical balance tests in the present study were sufficient to indicate large and conclusive changes in the probability of a fall event, three of the five clinical balance tests (ie, NBWT, TUG, FSST) had likelihood ratios that would imply small shifts in the probability of a fall given a positive or negative test (Table 4).55 For example, a score equal to or greater than 0.43 on the NBWT would suggest that a LLP user is three times more likely (ie, LR+ = 3.0) to be a faller than a nonfaller. Likelihood ratios for the NBWT, TUG, and FSST were accompanied by 95% CIs that did not overlap with 1.0 (ie, no change in probability) (Table 4), suggesting that the interpretation of an increase or decrease in the probability of being a faller based on a positive or negative test result can be made with a reasonable level of confidence.1 Although the increased likelihood of a fall associated with a positive test on the NBWT, TUG, and FSST relative to the BBS and ABC may be small, it may still be clinically important given the consequences of falls among LLP users.1,5,7,9,11 Although the LR were greater in the NBWT, TUG, and FSST, in many cases their CIs overlapped. However, the application of the indices derived in this study in a larger prospective study is needed to more definitively determine the impact of these differences.

Notably, fall risk assessment models that included multiple tests and factors associated with fall risk (ie, etiology and level of amputation)8,59 failed to improve likelihood ratios among LLP users compared to scores on the individual balance tests (Table 4). No additional information would therefore appear to be gained regarding the probability of being a faller among LLP users by combining scores from multiple balance tests with the other fall-related demographic or amputation information considered in this study. This suggests that amputation-related factors, including level or etiology of amputation, may provide less information regarding fall risk among established unilateral LLP users than has been historically considered. Additional research is required to verify this result in a prospective study, and to consider other fall-related demographic and amputation-specific information. As a result, clinicians may be best served by assessing fall risk in established unilateral LLP users by administering, scoring, and interpreting performance on a single balance test that possesses sufficient psychometric rigor, is practical for the given setting, and meets the application needs of the clinician (ie, discriminate, evaluate, or predict). Examples of such applications may include discriminating between fallers and nonfallers in an observational study, evaluating changes in balance ability pre-post therapy, and predicting fall risk to justify the prescription of prosthetic componentry.

Results of this study also indicate that the classification used to categorize participants as fallers or nonfallers can alter the validity indices of the studied clinical balance tests. Validity indices, such as the AUC, generally improved when fallers were classified as individuals reporting ≥2 falls. The observed increase in validity indices based on a more conservative faller classification (ie, ≥2 falls vs ≥1 fall)
is consistent with previous research. However, it seems important to note that limiting the classification of fallers to those with a history of multiple falls inherently creates a group of individuals with worse balance ability than a group that includes those who have fallen just once in the prior 12 months (Table 3). Adopting a more conservative ≥2 falls classification magnifies differences between fallers and nonfallers, increases the magnitude of validity indices, and improves each test’s ability to identify those at risk for additional falls. It does so, however, at the cost of being able to identify those at risk for a fall after experiencing only one fall in the prior 12 months, potentially overlooking those with modest balance impairments. The observed differences also suggest that studies reporting validity indices for clinical balance tests based on different faller classifications may not be directly comparable. For this reason, we recommend that investigators report cutoff scores and associated validity indices for both single and multiple fallers. Similarly, if a test’s validity indices are derived using one classification, test administrators should apply the same classification when applying the test clinically to assess fall risk. Cutoff scores for the FSST and TUG increased when fallers were defined as LLP users reporting multiple falls (ie, ≥2 falls), compared to users reporting one or more falls. This indicates that, when using these tests, administrators should use fall classification-specific cutoff scores (Table 4). In contrast, cutoff scores for the NBWT, BBS, and ABC did not change with how fallers were classified. This indicates that for these tests, administrators can use the same cutoff score irrespective of how they define a faller. Having a single cutoff score may simplify and reduce the burden of scoring and interpreting test performance, a reported restriction to the adoption of balance tests among clinicians. Whether cutoff scores could be established to differentiate LLP users experiencing no falls, a single, or multiple falls remains to be determined.

**Study Limitations**

A number of limitations with this study need to be considered when interpreting the results. First, the sample size, although consistent with similar studies, should be increased in future research. A larger sample would facilitate development of cutoff scores to discriminate nonfallers (ie, zero falls), from single fallers (ie, 1 fall), and multiple fallers (ie, ≥2 falls), as well as across score intervals for balance tests with a continuous scale (eg, the NBWT, TUG, and FSST). The study sample also consisted of a larger percentage of traumatic LLP users (ie, 62.5%) than is reported in the literature (ie, 17%-60.2%). Potentially limiting the generalization of study results. Although this is a challenge in most, but not all research involving LLP users, future efforts to include a larger proportion of dysvascular LLP users is required to increase generalization of study results. The mean age of our sample, 48.7 years, was slightly younger than that reported in larger, national studies of individuals with lower limb amputation (ie, mean age from 50 to 55), yet the range of ages studied, 24 to 70, is consistent with those prior studies. A follow-up study that focuses on older LLP users (eg, age ≥65) may be warranted to examine balance test cutoff scores in LLP users in that specific subpopulation (ie, Medicare-eligible individuals).

Falls were assessed retrospectively. This may underestimate fall frequency and lend itself to recall bias. A prospective study is needed to validate the cutoff scores and validity indices established in this cross-sectional study and establish a temporal relationship between balance and fall status.

Fall-related injuries were not recorded. Determining whether multiple versus nonmultiple fallers are more likely to suffer a fall-related injury, or if existing clinical tests can discriminate between or predict the probability of falls that result in injuries may be an important consideration in future studies.

Only a limited set of sociodemographic, health, and prosthetic-related factors were recorded and included in the multivariate models. Including other known risk factors for falls among lower limb prosthesis users such as strength, protective stepping, number of medications, and sense of vibration may improve model performance in future research.

Finally, psychometric properties, including the validity indices reported here, are population specific. The specific indices presented here therefore do not apply to other patient populations.

**Conclusion**

The primary objective of this study was to establish cutoff scores and associated validity indices for several clinical balance tests that may be administered to established unilateral LLP users. Given the limited options available to quantitatively assess fall risk among established unilateral LLP users, the proposed cutoff scores and associated likelihood ratios for the NBWT, TUG, and FSST provide clinicians with tools to reduce the uncertainty associated with estimating the probability of a fall among established unilateral LLP users. More studies establishing and assessing the accuracy of cutoff scores for diagnostic tests like these to predict outcomes among LLP users are urgently needed. Additional research to evaluate the relative reliability, utility, and prospective validity (ie, testing the cutoff scores and indices proposed here) of these tests is needed to facilitate their widespread adoption in clinical care.

**Supporting Information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.
References


**Inter-rater and test-retest reliability of performance-based clinical tests administered to established unilateral lower limb prosthesis users**

<table>
<thead>
<tr>
<th>Journal</th>
<th>Physical Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>PTJ-2019-0247.R1</td>
</tr>
<tr>
<td>Manuscript Category</td>
<td>Original Research - Measurement Property Evaluation</td>
</tr>
<tr>
<td>Section</td>
<td>Balance and Falls</td>
</tr>
<tr>
<td>Keywords</td>
<td>Amputation, Outcome Assessment (Health Care), Reliability, Accidental Falls, Reproducibility of Results</td>
</tr>
</tbody>
</table>
Running Head: Reliability of balance tests in amputees

Title: Inter-rater and test-retest reliability of performance-based clinical balance tests administered to established unilateral lower limb prosthesis users

Prior Publication: The work described in this manuscript has not been previously published, nor is it under consideration for publication elsewhere.

Financial Disclosure: The authors declare no competing interest.

Institutional Review: All study procedures were reviewed and approved by a XXXX institutional review board.

Clinical Trials Registration: Not applicable
Abstract (275/275 words)

Background: A major barrier to reducing falls among lower limb prosthesis (LLP) users has been an absence of statistical indices required for clinicians to select and interpret scores from performance-based clinical balance tests.

Objective: Derive estimates of reliability, measurement error, and minimal detectable change values in performance-based clinical balance tests administered to unilateral LLP users.

Design: Multi-site repeated-measures.

Methods: 60 unilateral LLP users were administered the Narrowing Beam Walking Test (NBWT), Timed Up and Go (TUG), Four Square Step Test (FSST), and 10 meter Walk Test (10mWT) on two occasions, three to nine days apart. Intraclass correlation coefficients (ICC) were calculated to assess inter-rater and test-retest reliability, while standard error of measurement (SEM) and minimal detectable change (MDC90) were derived to establish estimates of measurement error in individual scores or changes in score for each test.

Results: Inter-rater reliability ICCs (1,1) were high for all tests (i.e., ≥.98). With the exception of the TUG, studied tests had test-retest ICCs (2,1) that exceeded the minimum required threshold to be considered suitable for group- and individual-level applications (i.e., ICC ≥ .70 and ICC ≥ .90, respectively). SEM and MDC90 estimates varied between .07-.96 and .16-2.2 respectively.

Limitations: The study sample was skewed slightly, with a greater proportion of participants with traumatic and transtibial amputations than the general amputee population. Future research focusing on individuals with dysvascular amputations or in specific age categories (e.g., the U.S. Medicare population) may be required.

Conclusion: Along with published validity indices, these reliability, error, and change indices can help clinicians select balance tests suitable for LLP users. They can also help clinicians more effectively interpret test scores to make informed, evidence-based clinical decisions.
Key Words: amputation, accidental falls, outcome assessment, reliability, reproducibility of results.

3696/4000 words (introduction through conclusion)

Abbreviations:

10mWT = 10 meter Walk Test
ANOVA = Analysis of Variance
CCI = Charlson Comorbidity Index
COSMIN = COnsensus-based Standards for the selection of health Measurement INstruments
FSST = Four Square Step Test
ICC = intraclass correlation coefficient
LLP = lower limb prosthesis
MDC = minimal detectable change
MFCL = Medicare Functional Classification Level
NBWT = Narrowing Beam Walking Test
PLUS-M = Prosthetic Limb Users Survey of Mobility
REDCap = Research Electronic Data Capture
SCS = Socket Comfort Score
SD = standard deviation
SEM = standard error of measurement
STROBE: Strengthening The Reporting of Observational studies in Epidemiology
TUG = Timed Up and Go
Introduction

The use of valid and reliable instruments to document clinical outcomes is increasingly expected by employers and third-party payers in order to justify and evaluate rehabilitation services and interventions provided to lower limb prosthesis (LLP) users\textsuperscript{1-3}. Despite the demand for robust documentation, several barriers inhibit regular use of standardized outcome measures in clinical practice\textsuperscript{2-6}. Selection of suitable outcome measures for different clinical applications\textsuperscript{5,7-9}, and interpretation of test scores to guide clinical decision-making\textsuperscript{6,7} are frequently cited as challenges to routine use of outcome measures in clinical practice. These barriers may limit, among other things, regular assessment of balance ability and fall risk among LLP users, potentially contributing to the persistence of falls\textsuperscript{10-14}, and their related consequences\textsuperscript{15-17} among LLP users.

Indices of reproducibility (e.g., estimates of reliability, error, and detectable change), in the population of interest (i.e., LLP users)\textsuperscript{18-20}, are needed to formulate recommendations for performance-based clinical balance tests suited to the care of LLP users\textsuperscript{18,19,21}. Inter-rater and test-retest reliability, often quantified using intraclass correlation coefficients (ICC)\textsuperscript{22}, are estimates of the stability of measurement between raters and over a period of time when no changes in health are believed to have occurred\textsuperscript{19,23,24}. ICCs therefore provide test administrators with important information about the consistency of measurement(s) produced by a test within a given population. The magnitude of reliability estimates can also be used to guide clinicians in selecting performance-based tests for specific applications. To be recommended for group-level comparisons, a clinical test should generally demonstrate a test-retest ICC of .70 or greater\textsuperscript{25-28}. For applications that involve individual-level decisions, a test should have higher reliability, recommended as .90 or greater\textsuperscript{19,25,28}. Reliability estimates do not however help clinicians interpret test scores or provide the information needed to make evidence-based decisions for individual patients. The interpretation of individual test scores requires knowledge of the standard error of measurement (SEM) and minimal detectable change (MDC) of a test.
The SEM reflects the error associated with a single measurement. The SEM can therefore be used to estimate the precision of an individual’s score on a test, as it can be used to derive confidence intervals (CIs) around a score. The MDC is an estimate of the amount of change needed to exceed the error associated with two measurements on the same test within the same individual. An individual patient is considered to have changed only when the difference between the previous score and the current score exceeds the MDC of the test. The SEM and MDC can therefore help clinicians interpret test scores among individual patients by monitoring whether there is a real difference in their health or performance with respect to a reference value (i.e., using SEM), and/or whether their health or performance has changed between testing sessions (i.e., using MDC).

Few performance-based clinical balance tests with evidence of validity among LLP users have been assessed for evidence of reproducibility. In the absence of reliability indices like test-retest ICCs or SEM and MDC values, clinicians cannot select tests suitable for individual level decision-making, or interpret test scores of individual LLP users in clinical practice. Efforts to study the reproducibility of performance-based clinical balance tests administered to LLP users have been limited by several factors. First, sample sizes are generally small, and often below recommendations for establishing reliability indices. Second, evidence of reproducibility for some tests is dated and may not reflect best practices for assessing test-retest reliability. Lastly, study methodology (e.g., selection of raters and corresponding analysis) may limit the generalizability of study results to the broader population of LLP users and test administrators. Additionally, a number of common performance-based clinical balance tests (e.g., FSST, 10mWT, TUG) have not yet been assessed for inter-rater reliability in LLP users. Many performance-based tests also evaluate unique aspects of balance and walking (i.e., straight ahead walking, turns, balancing along beam, and multi-planar stepping). As such, it is important to evaluate each test and determine if it...
and its components are suitable for clinical practice. Finally, given the relative nature of test-retest reliability is relative to the sample of participants from which it is derived, and the evaluation of a test’s reliability is a continual process (i.e., tests are not reliable or unreliable, they exhibit more or less evidence of reliability), it is beneficial to accumulate additional evidence of reliability for a test, by comparing, contrasting, and confirming values across different samples of LLP users. There is therefore insufficient psychometric evidence to generate recommendations for the selection and interpretation of common performance-based clinical balance tests administered to LLP users.

The objective of this study was therefore to assess the reproducibility of performance-based clinical balance tests administered to established unilateral LLP users, deriving estimates of reliability indices to help clinicians and researchers select and interpret the studied tests. Specifically, the ICCs for inter-rater and test-retest reliability were assessed, from which the standard SEM and MDC of each of the studied tests were derived. We hypothesized that the studied tests (i.e., NBWT, FSST, TUG, and 10mWT) would demonstrate evidence of reliability matching or exceeding that required for group-level comparisons (i.e., ICC ≥ 0.7) and individual decision-making (i.e., ICC ≥ 0.9) in assessments of balance and fall risk.

**Methods**

**Study Design**

A two-site (XXXX and XXXX) repeated measures study was conducted between February 2018 and November 2018 to assess the reproducibility of the Narrowing Beam Walking Test (NBWT), Timed Up and Go (TUG), Four Square Step Test (FSST), and the 10 meter Walk Test (10mWT) when administered to LLP users. Best practices in conducting observational studies (i.e., STROBE: Strengthening The Reporting of Observational studies in Epidemiology), as well as the collection and
reporting of reliability data (i.e., COncensus-based Standards for the selection of health Measurement INstruments (COSMIN))\textsuperscript{19,20,37} were followed during the study. All data were stored and managed using a REDCap database hosted at XXXX\textsuperscript{41}. Study protocols were reviewed and approved by institutional review boards at each site. All individuals provided written consent prior to participation.

**Participants**

**Study flyers were used to recruit** LLP users from local prosthetic clinics in XXXX and XXXX. A convenience sampling method was used. Inclusion criteria were 18 years of age or older; unilateral amputation at the transtibial or transfemoral level; amputation due to trauma, dysvascular complications, tumor, or infection; use of a prosthesis for at least one year (i.e., established lower limb prosthesis users); use of a comfortable prosthesis to ambulate; able to walk 10 meters without an upper extremity mobility aid (e.g., cane); and able to read, write, and speak English. Participants were excluded from participation if they had an amputation of a second limb (e.g., arm or contralateral leg); contralateral complications (e.g., hip replacement, ulcers or infections), advanced neurologic disorder (e.g., Stroke, Multiple Sclerosis, Parkinson’s disease); or severe cardiopulmonary disease that would prevent them from completing study procedures.

An a priori sample size of 38 LLP users was estimated based on a desired 95% confidence interval with a width greater than or equal to .10 around an ICC of .80\textsuperscript{45}. Psychometric standards, however, recommend a minimum sample of 50 participants to estimate reliability statistics\textsuperscript{19}. Based on the more conservative estimate (i.e., recommended psychometric standards), and planning for modest attrition (i.e., 20%) between test and retest sessions\textsuperscript{46,47}, a sample of 60 LLP users was targeted.

**Procedures**
At their first visit, all participants were administered demographic and characterization measures, as well as four performance-based clinical balance tests. At the end of the first test session, each participant was scheduled a retest session three to nine days later. This test-retest interval was chosen to provide sufficient time to limit familiarity with the tests, and to avoid the possibility of changes in health or prosthetic fit. Participants wore their desired prosthesis and footwear, both of which were held constant between test sessions.

Measurements

Participant demographic and characterization measures

Sociodemographic information including age, height, weight, gender, race, and ethnicity was collected via self-report. Amputation characteristics (i.e., level, etiology, and time since amputation), prosthetic-related factors (i.e., components, and time since delivery), and mobility (i.e., Medicare Functional Classification Level (MFCL)\textsuperscript{48}, Prosthetic Limb Users Survey of Mobility (PLUS-M)\textsuperscript{49}, and typical prosthesis use/day) were collected via self-report, interview, and/or clinical inspection. The number of comorbid conditions was collected via self-report using the Charlson Comorbidity Index (CCI)\textsuperscript{50}. Comfort of the prosthesis worn by each participant was evaluated using the Socket Comfort Score (SCS), scored from 0 to 10\textsuperscript{51,52}. At retest, participants were asked whether they had experienced changes in their prosthesis and/or health since the first testing session. The SCS was re-administered at retest to confirm that participants’ scores were within the MDC\textsubscript{90} of the SCS in LLP users (i.e., ±2.82 points)\textsuperscript{52}.

Administration of performance-based clinical balance tests

The TUG\textsuperscript{40}, FSST\textsuperscript{41} 10mWT\textsuperscript{42}, and the NBWT\textsuperscript{39} were administered in a randomized order, and scored according to standardized instructions (Appendix 1). The same test sequence was used in both test sessions. Each balance test was selected because it could be performed in less than five minutes, has
been designed or used to assess fall risk in other populations, and has evidence of validity among LLP
users\textsuperscript{35,39,53,54}.

Standardized protocols were established to ensure fidelity in test administration and scoring between
study sites and raters. All raters were instructed in test procedures using a common set of training
documents. All raters practiced administration of the tests under the supervision of study (or site)
principal investigators. Test equipment was purchased or fabricated using explicit written instructions,
and verified with photographs and videos during a training session. Verbal instructions were
standardized to help ensure consistency in data collection across the two study sites\textsuperscript{55}. Further, no
communication was allowed between concurrent raters during testing to limit their awareness of
each other’s assessments, which could influence scoring\textsuperscript{56}. Finally, data collected on standardized paper
forms were double entered by research staff to minimize data-entry errors\textsuperscript{57}. Participants were
provided two minutes rest between tests, more if requested.

A small pool of qualified raters was consistently used at each study site. Specifically, across
study sites raters included two researchers (i.e., PhD), three PhD students (one certified prosthetist-
ortotist, one physical therapist, and one bioengineer), and three prosthetic and orthotic graduate
students. These raters were selected because of their qualifications, training, and knowledge of the
instruments under investigation. Further, these qualities reflected those of individuals working in a
clinical or research setting, thereby contributing to the generalizability of the subsequent reliability
indices\textsuperscript{20}. Testing during session two was performed by one of the raters from session one.

Data Analysis

Continuous variables were evaluated for departures from normality using the Shapiro-Wilk test\textsuperscript{58}.
Measures of central tendency and dispersion were calculated for each variable based on their
distribution. Paired 2-sided t-tests were used to test for differences in test scores between raters, and between test sessions (i.e., effect of time). A Bonferroni correction ($\alpha = .0125$) was used for multiple comparisons within the sample.

Inter-rater reliability of each clinical balance test was evaluated by calculating the ICC (1,1) from measurements taken simultaneously by two raters per participant$^{22,38,59}$. Each clinical balance test was also evaluated for **absolute** test-retest reliability by calculating the ICC (2,1) from measurements taken on two visits by the same rater$^{22,38}$. 95% confidence intervals (CI) were derived for each ICC using the F-distribution$^{60}$. Test-retest ICCs were compared to established thresholds in order to determine the suitability of each test for clinical and research applications (i.e., .70 or greater for group-level comparisons$^{25-28}$, and .90 or greater for individual-level decisions$^{19,25-28}$).

Standard error of measurement (SEM) estimates were calculated as $\text{SEM} = SD \times \sqrt{(1 - ICC)}$, where $SD$ is the pooled standard deviation of test scores across both test sessions from the same rater (i.e., rater A)$^{30}$, and ICC is the test-retest intraclass correlation coefficient$^{29,33}$. Alternatives sources of the standard deviation were explored (e.g., standard deviation from the first test session$^{33}$, and total sum of the squares from a one-way ANOVA$^{33,38,59}$), but neither resulted in an appreciable change in SEM values. MDC estimates were calculated as $\text{MDC}_{90} = 1.64 \times \sqrt{\text{SEM}^2} \times \sqrt{2}$ where 1.64 is the z-score for a 90% confidence interval$^{38}$.

**Results**

Sixty ambulatory unilateral LLP users with mean age of 55.1 years (SD = 21.3 years), 42 of which were male, were recruited and completed the study. Amputation, prosthesis, and mobility characteristics of the study sample are presented in Table 1. 48% of participants reported no co-morbid conditions, while 20%, 13%, and 10% reported one, two, or three or more co-morbid conditions (e.g., asthma,
diabetes, HIV/AIDS), respectively. Observed test scores (Supplement 1) for the FSST, TUG, and 10mWT were non-normally distributed (Shapiro-Wilk W=.798-.919, p≤.001). In contrast, NBWT scores were normally distributed (Shapiro-Wilk W=.957-.974, p≥.195). Despite departures from normality, parametric 2-sided t-tests were used to compare test scores between rater and across test sessions because of their robustness to withstand violations of normality in larger samples (i.e., n>30)\textsuperscript{61}.

**Inter-rater Reliability**

Two-sided paired t-tests revealed no statistically significant differences in mean test scores between raters for the TUG (t=1.609, p=.113), FSST (t=-.753, p=.455), NBWT (t=1.276, p=.207), or 10mWT (t=-.518, p=.607) (Table 2). Inter-rater reliability ICCs (model 1,1) were high (i.e., ≥.98) for all clinical tests (Table 2).

**Test-retest Reliability**

Retest sessions occurred at a mean of 6.7 days (SD = 2.3 day) after the initial testing session. No changes in health, prosthetic prescription, prosthesis function, or SCS (test session 1: 7.62 (SD=1.97), test session 2: 7.43 (SD=1.61), t=1.026, p=.309) were reported between test sessions by study participants, as a group or individually. Statistically significant differences in test scores were observed between test sessions for the TUG (t=2.968, p=.004), but not the 10mWT (t=2.315, p=.024), FSST (t=.821, p=.415), or NBWT (t=-1.821, p=.066) (Table 3). Test-retest ICCs (model 2,1) varied by test, ranging from .88 for the TUG, to .97 for the FSST (Table 3).

**Standard Error of Measurement and Minimal Detectable Change**

SEM and MDC\textsubscript{90} varied between .39 - .96, and .91s - 2.2s, respectively, among the time-based tests (i.e., 10mWT, TUG, and FSST) (Table 3). Estimates of SEM and MDC\textsubscript{90} for the NBWT, a test of normalized walking distance, were .07 and .16, respectively.
Discussion

The objectives of this study were to assess the reproducibility of performance-based clinical balance tests administered to established unilateral LLP users, and derive estimates of reliability indices to help clinicians and researchers select and interpret the studied tests. ICCs indicated that all studied tests are appropriate for group-level comparisons (i.e., ICC ≥ .70), while the FSST, 10mWT, and NBWT are also suitable for individual-level applications among established unilateral LLP users (i.e., ICC ≥ .90).

With the exception of the TUG, the results supported the study hypothesis that the studied tests (i.e., NBWT, FSST, TUG, and 10mWT) would demonstrate evidence of reliability matching or exceeding that required for group-level comparisons (i.e., ICC ≥ 0.7) and individual decision-making (i.e., ICC ≥ 0.9). Estimates of the SEM and MDC for the studied tests provide clinicians and researchers with indices to assess the precision of individual measurements (i.e., SEM), and to evaluate changes in measurements over time (i.e., MDC90). Our sample (n=60) was larger than previous reproducibility studies of performance-based clinical balance tests with LLP users (i.e., n=15-44)\(^{33,34}\), and exceeded the minimum threshold (i.e., n=50) recommended by COSMIN recommendations for “good” estimates of reliability indices\(^{37}\). The distribution of demographic (e.g., age), amputation (e.g., level, etiology, and time since), and activity characteristics (e.g., PLUS-M, MFCL) of participants in the present study were generally comparable to those reported in larger national studies of individuals with lower limb amputation (i.e., n=210 to 1568)\(^{52,62-65}\), although skewed slightly towards a higher proportion of traumatic and transtibial amputations. The overall similarity, however, with these samples suggests that the current results can be generalized to the broader population of established unilateral LLP users.

Reliability indices in the literature

The magnitude of the reliability indices (i.e., test-retest ICC, SEM, and MDC) for the studied tests varied with respect to previously reported values. Test-retest ICCs (Table 3) were consistent with those
previously reported for the 10mWT (i.e., .93-.98)\textsuperscript{66}, as well as the TUG, as described by Resnik et al., (2011) (i.e., .88 [95\% CI: .80-.94]), but not Clemens et al., (2018) (i.e., .98 [95\% CI: .97-.99]) or Jayakaran et al., (2011) (i.e., .95 [95\% CI: .85-.98]). The larger TUG test-retest ICC reported by Clemens et al., (2018) may be explained by their experimental design. Calculating the test-retest ICC using back-to-back trials of the TUG, rather than trials administered several days apart\textsuperscript{33,38}, excluded session-to-session variation within subjects, an important component of error when calculating test-retest ICCs\textsuperscript{38}. Excluding session-to-session variation likely increases the magnitude of test-retest ICCs and attenuates the corresponding indices of measurement error. The test-retest ICC for the TUG reported by Jayakaran et al., (2011) may have been affected by their atypical scoring (i.e., the mean of two trials rather than the conventional “best” of two trials)\textsuperscript{40}. Using the mean may have stabilized the score, reduced error, and increased reliability. To date, evidence of test-retest reliability for the FSST has been limited to other patient populations (e.g., Parkinson’s disease, multiple sclerosis, post-stroke)\textsuperscript{67}, but has yet to be reported among LLP users.

Notable departures from previously reported values of the standard error of measurement (SEM) were found for the studied tests. The SEM for the TUG in the present study (i.e., .96) was larger than one previously reported value, .55\textsuperscript{36}, and smaller than another, 1.6\textsuperscript{33}. Since the SEM is an arithmetic function of the test-retest ICC and the standard deviation of test scores from a study sample, the differences can be attributed to the notably larger test-retest ICC reported by Clemens et al., (2018), and the nearly two-fold increase in between-subjects variability reported by Resnik et al., (2011), respectively\textsuperscript{68}. To date, the SEM for the 10mWT or FSST among LLP users has not been reported.

Point estimates for the MDC\textsubscript{90} also differed from previously reported values\textsuperscript{33,36}. Given that MDC is a scalar product of the SEM, these differences can be explained by the earlier discussion of observed differences in SEM. Despite the disparities in MDC\textsubscript{90} point estimates in this study and those reported
previously, the 95% CIs of MDC\textsubscript{90} derived here overlapped with previously reported values. This implies that while diverse, there is some consistency with regard to what constitutes a true change in performance over time within each of the studied tests.

This study provides the first evidence of inter-rater reliability using ICCs for the TUG, FSST, and 10mWT in LLP users (Table 2). The reported inter-rater ICC of the TUG (i.e., .99 [95% CI: .99-1.0]) is consistent with what has been reported using a different approach (i.e., .96 with Pearson’s r\textsuperscript{35}).

Application and interpretation of reliability indices

The reliability indices reported here can serve to assist clinicians and researchers in selecting and interpreting performance-based clinical balance tests\textsuperscript{38}. ICCs provide an indication of how confident users should be in their measurements, namely, what percentage of the observed score variance is due to true score variance versus that attributable to error. All of the studied tests were found to have inter-rater and test-retest ICCs between .98-.99 and .88-.97 and respectively (Tables 2, 3). This suggests that 98% to 99% of the observed score variance between raters, and 88% to 97% of observed score variance between test sessions is due to true score variance, with the remaining observed score variance attributable to error (i.e., 1% to 12%). Trained users familiar with these tests can therefore be confident in the individual measurements made with these performance-based tests. Beyond establishing a level of confidence in an instrument, the magnitude of a test-retest ICC also suggests whether a test may be suitable for group-level comparisons (i.e., ICC \geq .70)\textsuperscript{25-28}, and/or individual-level applications (i.e., ICC \geq .90)\textsuperscript{19,25,28}. With the exception of the TUG, studied tests (i.e., FSST, NBWT, and 10mWT) had test-retest ICCs \geq .90, indicating that they are each suitable for group as well as individual-level applications. Because the TUG test-retest ICC (.88) is below, albeit slightly, the recommended threshold of 0.9, users should be aware that this test has greater error relative to the other performance tests studied here, and may therefore be better suited to group-level comparisons rather than making decisions about
individual patients. This does not however, suggest that the TUG lacks reliability altogether, simply that the observed level of reliability in this study warrants caution when using the TUG for individual applications like evaluating an individual LLP user over time.

An additional consideration when interpreting test-retest ICC values is the ICC model used to estimate reliability. Model selection, which is governed both by study design and whether systematic error is included in the calculation of the reliability coefficient, determines if reliability results can be generalized beyond the participants and raters included in the study. Consistent with most reproducibility studies conducted to-date, test-retest ICCs in the present study were calculated using ICC (2,1), which includes both random and systematic error. Model 2 permits the generalization of reliability results to other raters with similar participants or patients. The reliability results of the present study can therefore be generalized to other unilateral LLP users and raters.

The SEM and MDC of the studied tests provide clinicians with indices they can use to determine precision of their patients’ scores, or determine whether changes have occurred in their patients. The SEM provides an estimate of precision in a score. The confidence interval (CI) that is derived with the SEM can be used, for example, in cross-sectional applications to assess an individual’s score relative to a cutoff score (e.g., a threshold established to indicate those at risk for falls). In contrast, MDC provides an index for use in longitudinal applications. MDC can be used to assess whether change over time in an individual patient’s performance (i.e., repeated assessments) are above and beyond changes that may be attributable to measurement error. Any change in a subject’s score, either above or below the previous score, greater than the MDC is considered a real change. For example, if a LLP user recorded an initial NBWT score of .38, completed a 4-week balance-training program, and was then re-evaluated on the NBWT with a score of .55, the difference would exceed the reported MDC of .16, indicating that a real change in balance ability had been observed.
Limitations

The present study only included established unilateral LLP users who had been using a prosthesis for greater than or equal to 16 years on average. Additional research is required to determine whether these results extend to acute LLP users (i.e., < 1 year post amputation), bilateral LLP users, or other populations. Similarly, additional research focusing on individuals with dysvascular amputations or individuals in specific age categories (e.g., the U.S. Medicare population) may be required. A larger sample size would be required to determine whether reliability and validity indices vary between sub-groups of participants (i.e., LLP users with different levels of amputation, causes of amputation, or mobility levels).

The present study only evaluated the reproducibility of four performance-based clinical tests. Given the population-specific nature of validity and reliability indices, future research is required to evaluate other psychometric properties of these tests (e.g., minimally important clinical difference), as well as those of other performance-based tests that have proven useful in other patient populations (e.g., Functional Reach Test, Fullerton Advanced Balance Scale, Five-times Sit to Stand etc).

Significant session-to-session differences were observed in the TUG. Interestingly, the direction of these differences (i.e., time decreased, performance improved) differed from those observed in a previous study, which found differences of a similar magnitude but in the opposite direction (i.e., times increased, performance decreased). This difference may represent the natural variation in TUG performance among LLP users. It is also possible that differences in sample characteristics could contribute. For example, there appear to be differences in age (55 versus 66 years old), level of amputation (25% versus 50% transfemoral), the length of the test-retest period, and potentially
cause of amputation between the two studies. Additional research is however required to verify these between-session differences and offer potential explanations. Regardless, these “time effects” suggest that practice effects may persist between test sessions for the TUG when it is applied with LLP users\textsuperscript{38,72,73}. While not evaluated in the present study, within-session trial-to-trial practice effects may also exist among a host of performance-based clinical tests. These practice effects may limit not only the reliability of performance-based tests, but also their validity (i.e., diagnostic accuracy)\textsuperscript{73}. Previous research on the NBWT has shown that changes to administration and scoring procedures (i.e., providing additional practice trials) can accommodate practice effects and improve psychometric performance\textsuperscript{73}. It is possible that changes to TUG administration and scoring procedures may be able to resolve these issues\textsuperscript{22}. It may therefore be prudent to examine performance-based clinical balance tests’ administration and scoring procedures to determine if they are susceptible to practice effects or other sources of systematic error\textsuperscript{38}. If they are, modifications to administration and scoring procedures may prove effective in accommodating within and between session practice effects.

Conclusion

Results of this study provide valuable evidence of inter-rater and test-retest reliability, measurement error, and minimal detectable change for performance-based clinical balance tests among established unilateral LLP users. These reliability metrics provide clinicians with actionable evidence they can use to select tests for use with individual patients, interpret the resultant test scores, and effectively use the test results to make informed clinical decisions. All studied tests demonstrated sufficient reliability to be considered appropriate for group-level comparisons (i.e., ICC ≥ .70), while the FSST, 10mWT, and NBWT were also found to be suitable for individual-level applications among established unilateral LLP users (i.e., ICC ≥ .90). SEM and MDC estimates of the studied tests provide clinicians and researchers with the information necessary to assess differences or changes in balance...
ability and/or fall risk among individual LLP users. In conjunction with published validity indices,[53,74] clinicians and researchers can use the calculated values to better select these clinical tests for LLP users, and more effectively interpret the resulting scores to make informed, evidence-based clinical decisions.

References


Table 1. Amputation, prosthetic, and activity characteristics of the study sample

<table>
<thead>
<tr>
<th>Amputation Characteristics</th>
<th>Prosthetic Characteristics</th>
<th>Activity Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Etiology</td>
<td>Time Since Amputation (years)</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Transtibial (n=45)</td>
<td>Trauma (n=26)</td>
<td>16.0 (14.8)</td>
</tr>
<tr>
<td></td>
<td>Infection (n=13)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dysvascular (n=11)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tumor (n=9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other (n=1)</td>
<td></td>
</tr>
<tr>
<td>Transfemoral (n=15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence Interval, IQR: Interquartile range, MFCL: Medicare Functional Classification Level (K-level), SACH: solid ankle cushioned heel
<table>
<thead>
<tr>
<th>Test</th>
<th>Rater A Session 1 Score Mean (SD)</th>
<th>Rater B Session 1 Score Mean (SD)</th>
<th>Score Difference Mean (SD)</th>
<th>ICC (1,1) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (score, /1.00)</td>
<td>.42 (0.21)</td>
<td>.41 (0.22)</td>
<td>0.01 (0.03)</td>
<td>.99 (.98-.99)</td>
</tr>
<tr>
<td>TUG (time, s)</td>
<td>10.4 (2.76)</td>
<td>10.4 (2.79)</td>
<td>0.05 (0.27)</td>
<td>.99 (.99-1.0)</td>
</tr>
<tr>
<td>FSST (time, s)</td>
<td>10.4 (5.27)</td>
<td>10.4 (5.21)</td>
<td>-0.03 (-0.37)</td>
<td>.99 (.99-1.0)</td>
</tr>
<tr>
<td>10MWT (time, s)</td>
<td>8.74 (1.96)</td>
<td>8.76 (1.96)</td>
<td>-0.02 (-0.30)</td>
<td>.99 (.98-.99)</td>
</tr>
</tbody>
</table>

NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement
Table 3. Test-retest reliability of performance-based clinical balance tests in people with lower-limb loss

<table>
<thead>
<tr>
<th>Test</th>
<th>Session 1 Score Mean (SD)</th>
<th>Session 2 Score Mean (SD)</th>
<th>Score Change Mean (SD)</th>
<th>ICC (2,1) (95% CI)</th>
<th>SEM</th>
<th>MDC&lt;sub&gt;90&lt;/sub&gt; (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (score, /1.00)</td>
<td>.42 (.21)</td>
<td>.44 (.24)</td>
<td>.02 (.10)</td>
<td>.90 (.83-.94)</td>
<td>.07</td>
<td>.16 (.10-.20)</td>
</tr>
<tr>
<td>TUG (time, s)</td>
<td>10.4 (2.76)</td>
<td>9.91 (2.75)</td>
<td>.50 (1.30)*</td>
<td>.88 (.78-.93)</td>
<td>.96</td>
<td>2.2 (1.5-2.9)</td>
</tr>
<tr>
<td>FSST (time, s)</td>
<td>10.4 (5.27)</td>
<td>10.3 (4.83)</td>
<td>.14 (1.32)</td>
<td>.97 (.94-.98)</td>
<td>.87</td>
<td>2.0 (.79-3.5)</td>
</tr>
<tr>
<td>10MWT (time, s)</td>
<td>8.74 (1.96)</td>
<td>8.58 (2.00)</td>
<td>.16 (.54)</td>
<td>.96 (.93-.98)</td>
<td>.39</td>
<td>.91 (.40-1.4)</td>
</tr>
</tbody>
</table>

* Test score significantly different between test sessions (2-sided paired t-test) p=.004

NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement, MDC<sub>90</sub>: Minimal Detectable Change
Timed Up and Go (TUG) Protocol

**Explanation to participant:** The goal of this test is to rise from the chair, walk around the cone, walk back to the chair and sit down again. You should walk at your normal, comfortable pace. I will time you while you perform the test.

**Demonstration:** Demonstrate the test one time.

**Test instructions:** Begin the test sitting with your back against the back of the chair and your arms resting on the armrests. When I say go, please stand up and walk around the cone, walk back to the chair, and sit down again. Please walk at your normal, comfortable pace.

**Practice:** Administer one practice trial. Do not time the practice trial.

**Administration:** Administer the test 2 times. Begin timing when you say go. Stop timing when the participants’ buttocks touch the chair.

**Scoring:** Select faster of the two timed trials as the TUG score.
Four Square Step Test (FSST) Protocol

**Explanation to participant:** The goal of this test is to step over the canes in a specific sequence as quickly as possible. I will time you while you perform this test.

**Demonstration:** Demonstrate the test one time. Demonstrate starting in square 1 and stepping in squares 2, 3, 4, 1, 4, 3, 2, and 1.

**Test instructions:** When I say go, please step in the sequent I demonstrated. Try to complete the sequence as fast as possible without touching the canes. Both feet must make contact with the floor in each square. Face forward during the entire sequence.

**Practice:** Administer one practice trial. Do not time the practice trial.

**Administration:** Administer the test 2 times. Begin timing when the first foot contacts square 2. Stop timing when the last foot contacts square 1. Repeat the trial if the participant does not complete the sequence, loses balance, or contact a cane.

**Scoring:** Select the faster of the two timed trials as the FSST score.
Narrowing Beam Walking Test (NBWT) Protocol

Explanation to participant: The goal of this test is to walk as far as possible along the beam. Speed is not being evaluated. Begin the test by standing with one foot on the wide end of the beam and the other foot on the ground to the side. You may choose which foot to put on the beam and which to put on the ground. Please cross both your arms across your chest.

Demonstration: Demonstrate the test one time.

Test instructions: When I say go, please walk along the beam as far as you can. Please walk at a comfortable speed. Remember to keep your arms crossed over your chest as you walk. Once you move your arms away from your body or step off the beam, I will ask you to stop.

Practice: Do not administer a practice trial.

Administration: Administer the test 5 times. Stop the trial when a participant: walks the length of the beam, steps off the beam, or moves their arms away from their body.

Scoring: Average the distances walked during trials 3 through 5, and divide by 22. This normalized distance (i.e., 0-1) is recorded as the NBWT score. The average distance is divided by 22 and not 24 (the total length of the beam), because participants begin the test with one foot on the beam and must walk past the 0.61m (2.0ft) mark to receive a score above 0.0.
10 Meter Walk Test (10MWT) Protocol

**Explanation to participant:** The goal of this test is to walk a short distance at your preferred comfortable walking speed. I will time you while you perform this test.

**Demonstration:** Demonstrate the test one time.

**Test instructions:** When I say ‘go,’ please walk at your normal, comfortable pace until I say stop.

**Practice:** Administer one practice trial.

**Administration:** Prepare a 14-meter walkway in a hallway or other unobstructed area. Place lines at 0, 2, 12, and 14 meters. Have the participant start in a standing position on the 0-meter line. Inform them that, on the word “go,” they are to walk at a comfortable speed until you say “stop.” Begin timing when the participant crosses the 2-meter line. Stop timing when the participant crosses the 12-meter line. Inform the participant to stop when they cross the 14-meter line. Repeat the test two times.

**Scoring:** Select the faster of the two timed trials as the TUG score.
Running Head: Reliability of balance tests in amputees

Title: Inter-rater and test-retest reliability of performance-based clinical balance tests administered to established unilateral lower limb prosthesis users

Authors Andrew Sawers, CPO, PhD¹, Janis Kim, MPT¹, Geoff Balkman², Brian Hafner, PhD²

Author Contribution:

Study Concept and Design: A. Sawers, B. Hafner

Acquisition of Data: A. Sawers, J. Kim, G. Balkman, B. Hafner

Analysis and Interpretation of Data: A. Sawers, B. Hafner

Drafting of Manuscript: A. Sawers

Critical Revision of Manuscript of Important Intellectual Content: B. Hafner, J. Kim, G. Balkman

Statistical Analysis: A. Sawers

Study Supervision: A. Sawers

Institutions and Affiliations:

¹ Department of Kinesiology, University of Illinois at Chicago, Chicago, IL 60612

² Department of Rehabilitation Medicine, University of Washington, Seattle, WA 98195

Prior Publication:

The work described in this manuscript has not been previously published, nor is it under consideration for publication elsewhere.
**Funding Support:** Research reported in this publication was supported by the Department of Defense (DoD) under award number W81XWH-17-1-0547 (BH, AS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the DoD.

**Financial Disclosure:** The authors declare no competing interest.

**Institutional Review:** All study procedures were reviewed and approved by a University of Illinois at Chicago institutional review board.

**Corresponding Author:**
Andrew Sawers, CPO, PhD
Department of Kinesiology
University of Illinois at Chicago
1919 West Taylor Street, Rm. 646
Chicago, IL, 60612
Phone: 312-996-1427
Fax: 312-413-0319
Email: asawers@uic.edu

**Clinical Trials Registration:** Not applicable
Abstract (275/275 words)

**Background**: A major barrier to reducing falls among lower limb prosthesis (LLP) users has been an absence of statistical indices required for clinicians to select and interpret scores from performance-based clinical balance tests.

**Objective**: Derive estimates of reliability, measurement error, and minimal detectable change values in performance-based clinical balance tests administered to unilateral LLP users.

**Design**: Multi-site repeated-measures.

**Methods**: 60 unilateral LLP users were administered the Narrowing Beam Walking Test (NBWT), Timed Up and Go (TUG), Four Square Step Test (FSST), and 10 meter Walk Test (10mWT) on two occasions, three to nine days apart. Intraclass correlation coefficients (ICC) were calculated to assess inter-rater and test-retest reliability, while standard error of measurement (SEM) and minimal detectable change (MDC$_{90}$) were derived to establish estimates of measurement error in individual scores or changes in score for each test.

**Results**: Inter-rater reliability ICCs (1,1) were high for all tests (i.e., ≥.98). With the exception of the TUG, studied tests had test-retest ICCs (2,1) that exceeded the minimum required threshold to be considered suitable for group- and individual-level applications (i.e., ICC ≥ .70 and ICC ≥ .90, respectively). SEM and MDC$_{90}$ estimates varied between .07-.96 and .16-2.2 respectively.

**Limitations**: The study sample was skewed slightly, with a greater proportion of participants with traumatic and transtibial amputations than the general amputee population. Future research focusing on individuals with dysvascular amputations or in specific age categories (e.g., the U.S. Medicare population) may be required.

**Conclusion**: Along with published validity indices, these reliability, error, and change indices can help clinicians select balance tests suitable for LLP users. They can also help clinicians more effectively interpret test scores to make informed, evidence-based clinical decisions.
**Key Words:** amputation, accidental falls, outcome assessment, reliability, reproducibility of results.

3696/4000 words (introduction through conclusion)

**Abbreviations:**

- 10mWT = 10 meter Walk Test
- ANOVA = Analysis of Variance
- CCI = Charlson Comorbidity Index
- COSMIN = COnsensus-based Standards for the selection of health Measurement INstruments
- FSST = Four Square Step Test
- ICC = intraclass correlation coefficient
- LLP = lower limb prosthesis
- MDC = minimal detectable change
- MFCL = Medicare Functional Classification Level
- NBWT = Narrowing Beam Walking Test
- PLUS-M = Prosthetic Limb Users Survey of Mobility
- REDCap = Research Electronic Data Capture
- SCS = Socket Comfort Score
- SD = standard deviation
- SEM = standard error of measurement
- STROBE: Strengthening The Reporting of Observational studies in Epidemiology
- TUG = Timed Up and Go
Introduction

The use of valid and reliable instruments to document clinical outcomes is increasingly expected by employers and third-party payers in order to justify and evaluate rehabilitation services and interventions provided to lower limb prosthesis (LLP) users\(^1-3\). Despite the demand for robust documentation, several barriers inhibit regular use of standardized outcome measures in clinical practice\(^2-6\). Selection of suitable outcome measures for different clinical applications\(^5,7-9\), and interpretation of test scores to guide clinical decision-making\(^6,7\) are frequently cited as challenges to routine use of outcome measures in clinical practice. These barriers may limit, among other things, regular assessment of balance ability and fall risk among LLP users, potentially contributing to the persistence of falls\(^10-14\), and their related consequences\(^15-17\) among LLP users.

Indices of reproducibility (e.g., estimates of reliability, error, and detectable change), in the population of interest (i.e., LLP users)\(^18-20\), are needed to formulate recommendations for performance-based clinical balance tests suited to the care of LLP users\(^18,19,21\). Inter-rater and test-retest reliability, often quantified using intraclass correlation coefficients (ICC)\(^22\), are estimates of the stability of measurement between raters and over a period of time when no changes in health are believed to have occurred\(^19,23,24\). ICCs therefore provide test administrators with important information about the consistency of measurement(s) produced by a test within a given population. The magnitude of reliability estimates can also be used to guide clinicians in selecting performance-based tests for specific applications. To be recommended for group-level comparisons, a clinical test should generally demonstrate a test-retest ICC of .70 or greater\(^25-28\). For applications that involve individual-level decisions, a test should have higher reliability, recommended as .90 or greater\(^19,25,28\). Reliability estimates do not however help clinicians interpret test scores or provide the information needed to make evidence-based decisions for individual patients. The interpretation of individual test scores requires knowledge of the standard error of measurement (SEM) and minimal detectable change (MDC) of a test.
The SEM reflects the error associated with a single measurement\(^ {19}\). The SEM can therefore be used to estimate the precision of an individual’s score on a test\(^ {19,29,30}\), as it can be used to derive confidence intervals (CIs) around a score. The MDC is an estimate of the amount of change needed to exceed the error associated with two measurements on the same test within the same individual\(^ {31}\). An individual patient is considered to have changed only when the difference between the previous score and the current score exceeds the MDC of the test\(^ {32}\). The SEM and MDC can therefore help clinicians interpret test scores among individual patients by monitoring whether there is a real difference in their health or performance with respect to a reference value (i.e., using SEM), and/or whether their health or performance has changed between testing sessions (i.e., using MDC)\(^ {19}\).

Few performance-based clinical balance tests with evidence of validity among LLP users have been assessed for evidence of reproducibility\(^ {33-36}\). In the absence of reliability indices like test-retest ICCs or SEM and MDC values, clinicians cannot select tests suitable for individual level decision-making, or interpret test scores of individual LLP users in clinical practice. Efforts to study the reproducibility of performance-based clinical balance tests administered to LLP users have been limited by several factors. First, sample sizes are generally small \(^ {33-35}\), and often below recommendations for establishing reliability indices\(^ {37}\). Second, evidence of reproducibility for some tests is dated and may not reflect best practices for assessing test-retest reliability\(^ {35}\). Lastly, study methodology (e.g., selection of raters and corresponding analysis) may limit the generalizability of study results to the broader population of LLP users and test administrators\(^ {34,36}\). Additionally, a number of common performance-based clinical balance tests (e.g., FSST, 10mWT, TUG) have not yet been assessed for inter-rater reliability in LLP users. Many performance-based tests also evaluate unique aspects of balance and walking (i.e., straight ahead walking (10MWT), turns (TUG), balancing along beam (NBWT), and multi-planar stepping (FSST)). As such, it is important to evaluate each test and determine if it
and its components are suitable for clinical practice. Finally, given the relative nature of test-retest reliability is relative to the sample of participants from which it is derived\(^3\), and the evaluation of a test’s reliability is a continual process (i.e., tests are not reliable or unreliable, they exhibit more or less evidence of reliability), it is beneficial to accumulate additional evidence of reliability for a test, by comparing, contrasting, and confirming values across different samples of LLP users. There is therefore insufficient psychometric evidence to generate recommendations for the selection and interpretation of common performance-based clinical balance tests administered to LLP users.

The objective of this study was therefore to assess the reproducibility of performance-based clinical balance tests administered to established unilateral LLP users, deriving estimates of reliability indices to help clinicians and researchers select and interpret the studied tests. Specifically, the ICCs for inter-rater and test-retest reliability were assessed, from which the standard SEM and MDC of each of the studied tests were derived. We hypothesized that the studied tests (i.e., NBWT, FSST, TUG, and 10mWT) would demonstrate evidence of reliability matching or exceeding that required for group-level comparisons (i.e., ICC ≥ 0.7) and individual decision-making (i.e., ICC ≥ 0.9) in assessments of balance and fall risk.

**Methods**

**Study Design**

A two-site (Chicago and Seattle) repeated measures study was conducted between February 2018 and November 2018 to assess the reproducibility of the Narrowing Beam Walking Test (NBWT)\(^3\), Timed Up and Go (TUG)\(^4\), Four Square Step Test (FSST)\(^5\), and the 10 meter Walk Test (10mWT)\(^6\) when administered to LLP users. Best practices in conducting observational studies (i.e., STROBE: Strengthening The Reporting of Observational studies in Epidemiology\(^7\)), as well as the collection and
reporting of reliability data (i.e., COConsensus-based Standards for the selection of health Measurement INstruments (COSMIN))\textsuperscript{19,20,37} were followed during the study. All data were stored and managed using a REDCap database hosted at the University of Illinois at Chicago\textsuperscript{44}. Study protocols were reviewed and approved by institutional review boards at each site. All individuals provided written consent prior to participation.

Participants

Study flyers were used to recruit LLP users from local prosthetic clinics in Chicago and Seattle. A convenience sampling method was used. Inclusion criteria were 18 years of age or older; unilateral amputation at the transtibial or transfemoral level; amputation due to trauma, dysvascular complications, tumor, or infection; use of a prosthesis for at least one year (i.e., established lower limb prosthesis users); use of a comfortable prosthesis to ambulate; able to walk 10 meters without an upper extremity mobility aid (e.g., cane); and able to read, write, and speak English. Participants were excluded from participation if they had an amputation of a second limb (e.g., arm or contralateral leg); contralateral complications (e.g., hip replacement, ulcers or infections), advanced neurologic disorder (e.g., Stroke, Multiple Sclerosis, Parkinson’s disease); or severe cardiopulmonary disease that would prevent them from completing study procedures.

An a priori sample size of 38 LLP users was estimated based on a desired 95\% confidence interval with a width greater than or equal to .10 around an ICC of .80\textsuperscript{45}. Psychometric standards, however, recommend a minimum sample of 50 participants to estimate reliability statistics\textsuperscript{19}. Based on the more conservative estimate (i.e., recommended psychometric standards), and planning for modest attrition (i.e., 20\%) between test and retest sessions\textsuperscript{46,47}, a sample of 60 LLP users was targeted.

Procedures
At their first visit, all participants were administered demographic and characterization measures, as well as four performance-based clinical balance tests. At the end of the first test session, each participant was scheduled a retest session three to nine days later. This test-retest interval was chosen to provide sufficient time to limit familiarity with the tests, and to avoid the possibility of changes in health or prosthetic fit. Participants wore their desired prosthesis and footwear, both of which were held constant between test sessions.

**Measurements**

**Participant demographic and characterization measures**

Sociodemographic information including age, height, weight, gender, race, and ethnicity was collected via self-report. Amputation characteristics (i.e., level, etiology, and time since amputation), prosthetic-related factors (i.e., components, and time since delivery), and mobility (i.e., Medicare Functional Classification Level (MFCL)\(^{48}\), Prosthetic Limb Users Survey of Mobility (PLUS-M)\(^{49}\), and typical prosthesis use/day) were collected via self-report, interview, and/or clinical inspection. The number of comorbid conditions was collected via self-report using the Charlson Comorbidity Index (CCI)\(^{50}\). Comfort of the prosthesis worn by each participant was evaluated using the Socket Comfort Score (SCS), scored from 0 to 10\(^{51,52}\). At retest, participants were asked whether they had experienced changes in their prosthesis and/or health since the first testing session. The SCS was re-administered at retest to confirm that participants’ scores were within the MDC\(_{90}\) of the SCS in LLP users (i.e., ±2.82 points)\(^{52}\).

**Administration of performance-based clinical balance tests**

The TUG\(^{40}\), FSST\(^{41}\) 10mWT\(^{42}\), and the NBWT\(^{39}\) were administered in a randomized order, and scored according to standardized instructions (Appendix 1). The same test sequence was used in both test sessions. Each balance test was selected because it could be performed in less than five minutes, has
been designed or used to assess fall risk in other populations, and has evidence of validity among LLP
users\textsuperscript{35,39,53,54}.

Standardized protocols were established to ensure fidelity in test administration and scoring between
study sites and raters. All raters were instructed in test procedures using a common set of training
documents. All raters practiced administration of the tests under the supervision of study (or site)
principal investigators. Test equipment was purchased or fabricated using explicit written instructions,
and verified with photographs and videos during a training session. Verbal instructions were
standardized to help ensure consistency in data collection across the two study sites\textsuperscript{55}. Further, no
communication was allowed between concurrent raters during testing to limit their awareness of
each other’s assessments, which could influence scoring\textsuperscript{56}. Finally, data collected on standardized paper
forms were double entered by research staff to minimize data-entry errors\textsuperscript{57}. Participants were
provided two minutes rest between tests, more if requested.

A small pool of qualified raters was consistently used at each study site. Specifically, across
study sites raters included two researchers (i.e., PhD), three PhD students (one certified prosthetist-
orthotist, one physical therapist, and one bioengineer), and three prosthetic and orthotic graduate
students. These raters were selected because of their qualifications, training, and knowledge of the
instruments under investigation. Further, these qualities reflected those of individuals working in a
clinical or research setting, thereby contributing to the generalizability of the subsequent reliability
indices\textsuperscript{20}. Testing during session two was performed by one of the raters from session one.

Data Analysis

Continuous variables were evaluated for departures from normality using the Shapiro-Wilk test\textsuperscript{58}.

Measures of central tendency and dispersion were calculated for each variable based on their
distribution. Paired 2-sided t-tests were used to test for differences in test scores between raters, and between test sessions (i.e., effect of time). A Bonferroni correction ($\alpha = 0.0125$) was used for multiple comparisons within the sample.

Inter-rater reliability of each clinical balance test was evaluated by calculating the ICC (1,1) from measurements taken simultaneously by two raters per participant$^{22,38,59}$. Each clinical balance test was also evaluated for absolute test-retest reliability by calculating the ICC (2,1) from measurements taken on two visits by the same rater$^{22,38}$. 95% confidence intervals (CI) were derived for each ICC using the F-distribution$^{60}$. Test-retest ICCs were compared to established thresholds in order to determine the suitability of each test for clinical and research applications (i.e., .70 or greater for group-level comparisons$^{25-28}$, and .90 or greater for individual-level decisions$^{19,25-28}$).

Standard error of measurement (SEM) estimates were calculated as $SEM = SD \times \sqrt{1 - ICC}$, where SD is the pooled standard deviation of test scores across both test sessions from the same rater (i.e., rater A)$^{30}$, and ICC is the test-retest intraclass correlation coefficient$^{29,33}$. Alternatives sources of the standard deviation were explored (e.g., standard deviation from the first test session$^{33}$, and total sum of the squares from a one-way ANOVA$^{33,38,59}$), but neither resulted in an appreciable change in SEM values. MDC estimates were calculated as $MDC_{90} = 1.64 \times \sqrt{SEM \times \sqrt{2}}$ where 1.64 is the z-score for a 90% confidence interval$^{38}$.

**Results**

Sixty ambulatory unilateral LLP users with mean age of 55.1 years (SD = 21.3 years), 42 of which were male, were recruited and completed the study. Amputation, prosthesis, and mobility characteristics of the study sample are presented in Table 1. 48% of participants reported no co-morbid conditions, while 20%, 13%, and 10% reported one, two, or three or more co-morbid conditions (e.g., asthma,
diabetes, HIV/AIDS), respectively. Observed test scores (Supplement 1) for the FSST, TUG, and 10mWT were non-normally distributed (Shapiro-Wilk W=.798-.919, p≤.001). In contrast, NBWT scores were normally distributed (Shapiro-Wilk W=.957-.974, p≥.195). Despite departures from normality, parametric 2-sided t-tests were used to compare test scores between rater and across test sessions because of their robustness to withstand violations of normality in larger samples (i.e., n>30)\textsuperscript{61}.

**Inter-rater Reliability**

Two-sided paired t-tests revealed no statistically significant differences in mean test scores between raters for the TUG (t=1.609, p=.113), FSST (t=-.753, p=.455), NBWT (t=1.276, p=.207), or 10mWT (t=-.518, p=.607) (Table 2). Inter-rater reliability ICCs (model 1,1) were high (i.e., ≥.98) for all clinical tests (Table 2).

**Test-retest Reliability**

Retest sessions occurred at a mean of 6.7 days (SD = 2.3 day) after the initial testing session. No changes in health, prosthetic prescription, prosthesis function, or SCS (test session 1: 7.62 (SD=1.97), test session 2: 7.43 (SD=1.61), t=1.026, p=.309) were reported between test sessions by study participants, as a group or individually. Statistically significant differences in test scores were observed between test sessions for the TUG (t=2.968, p=.004), but not the 10mWT (t=2.315, p=.024), FSST (t=.821, p=.415), or NBWT (t=-1.821, p=.066) (Table 3). Test-retest ICCs (model 2,1) varied by test, ranging from .88 for the TUG, to .97 for the FSST (Table 3).

**Standard Error of Measurement and Minimal Detectable Change**

SEM and MDC\textsubscript{90} varied between .39 - .96, and .91s - 2.2s, respectively, among the time-based tests (i.e., 10mWT, TUG, and FSST) (Table 3). Estimates of SEM and MDC\textsubscript{90} for the NBWT, a test of normalized walking distance, were .07 and .16, respectively.
Discussion

The objectives of this study were to assess the reproducibility of performance-based clinical balance tests administered to established unilateral LLP users, and derive estimates of reliability indices to help clinicians and researchers select and interpret the studied tests. ICCs indicated that all studied tests are appropriate for group-level comparisons (i.e., ICC ≥ .70), while the FSST, 10mWT, and NBWT are also suitable for individual-level applications among established unilateral LLP users (i.e., ICC ≥ .90).

With the exception of the TUG, the results supported the study hypothesis that the studied tests (i.e., NBWT, FSST, TUG, and 10mWT) would demonstrate evidence of reliability matching or exceeding that required for group-level comparisons (i.e., ICC ≥ 0.7) and individual decision-making (i.e., ICC ≥ 0.9). Estimates of the SEM and MDC for the studied tests provide clinicians and researchers with indices to assess the precision of individual measurements (i.e., SEM), and to evaluate changes in measurements over time (i.e., MDC90). Our sample (n=60) was larger than previous reproducibility studies of performance-based clinical balance tests with LLP users (i.e., n=15-44)33,34, and exceeded the minimum threshold (i.e., n=50) recommended by COSMIN recommendations for “good” estimates of reliability indices37. The distribution of demographic (e.g., age), amputation (e.g., level, etiology, and time since), and activity characteristics (e.g., PLUS-M, MFCL) of participants in the present study were generally comparable to those reported in larger national studies of individuals with lower limb amputation (i.e., n=210 to 1568)52,62-65, although skewed slightly towards a higher proportion of traumatic and transtibial amputations. The overall similarity, however, with these samples suggests that the current results can be generalized to the broader population of established unilateral LLP users.

Reliability indices in the literature

The magnitude of the reliability indices (i.e., test-retest ICC, SEM, and MDC) for the studied tests varied with respect to previously reported values. Test-retest ICCs (Table 3) were consistent with those...
previously reported for the 10mWT (i.e., .93-.98)\(^66\), as well as the TUG, as described by Resnik et al., (2011) (i.e., .88 [95% CI: .80-.94]), but not Clemens et al., (2018) (i.e., .98 [95% CI: .97-.99]) or Jayakaran et al., (2011) (i.e., .95 [95% CI: .85-.98]). The larger TUG test-retest ICC reported by Clemens et al., (2018) may be explained by their experimental design. Calculating the test-retest ICC using back-to-back trials of the TUG, rather than trials administered several days apart\(^33,38\), excluded session-to-session variation within subjects, an important component of error when calculating test-retest ICCs\(^38\). Excluding session-to-session variation likely increases the magnitude of test-retest ICCs and attenuates the corresponding indices of measurement error. The test-retest ICC for the TUG reported by Jayakaran et al., (2011) may have been affected by their atypical scoring (i.e., the mean of two trials rather than the conventional “best” of two trials)\(^40\). Using the mean may have stabilized the score, reduced error, and increased reliability. To date, evidence of test-retest reliability for the FSST has been limited to other patient populations (e.g., Parkinson’s disease, multiple sclerosis, post-stroke)\(^67\), but has yet to be reported among LLP users.

Notable departures from previously reported values of the standard error of measurement (SEM) were found for the studied tests. The SEM for the TUG in the present study (i.e., .96) was larger than one previously reported value, .55\(^36\), and smaller than another, 1.6\(^33\). Since the SEM is an arithmetic function of the test-retest ICC and the standard deviation of test scores from a study sample, the differences can be attributed to the notably larger test-retest ICC reported by Clemens et al., (2018), and the nearly two-fold increase in between-subjects variability reported by Resnik et al., (2011), respectively\(^68\). To date, the SEM for the 10mWT or FSST among LLP users has not been reported.

Point estimates for the MDC\(_{90}\) also differed from previously reported values\(^33,36\). Given that MDC is a scalar product of the SEM, these differences can be explained by the earlier discussion of observed differences in SEM. Despite the disparities in MDC\(_{90}\) point estimates in this study and those reported
previously, the 95% CIs of MDC$_{90}$ derived here overlapped with previously reported values. This implies that while diverse, there is some consistency with regard to what constitutes a true change in performance over time within each of the studied tests.

This study provides the first evidence of inter-rater reliability using ICCs for the TUG, FSST, and 10mWT in LLP users (Table 2). The reported inter-rater ICC of the TUG (i.e., .99 [95% CI: .99-1.0]) is consistent with what has been reported using a different approach (i.e., .96 with Pearson’s r)$^{35}$.

**Application and interpretation of reliability indices**

The reliability indices reported here can serve to assist clinicians and researchers in selecting and interpreting performance-based clinical balance tests$^{38}$. ICCs provide an indication of how confident users should be in their measurements, namely, what percentage of the observed score variance is due to true score variance versus that attributable to error. All of the studied tests were found to have inter-rater and test-retest ICCs between .98-.99 and .88-.97 and respectively (Tables 2, 3). This suggests that 98% to 99% of the observed score variance between raters, and 88% to 97% of observed score variance between test sessions is due to true score variance, with the remaining observed score variance attributable to error (i.e., 1% to 12%). Trained users familiar with these tests can therefore be confident in the individual measurements made with these performance-based tests. Beyond establishing a level of confidence in an instrument, the magnitude of a test-retest ICC also suggests whether a test may be suitable for group-level comparisons (i.e., ICC ≥ .70)$^{25-28}$, and/or individual-level applications (i.e., ICC ≥ .90)$^{19,25,28}$. With the exception of the TUG, studied tests (i.e., FSST, NBWT, and 10mWT) had test-retest ICCs ≥ .90, indicating that they are each suitable for group as well as individual-level applications.

Because the TUG test-retest ICC (.88) is below, albeit slightly, the recommended threshold of 0.9, users should be aware that this test has greater error relative to the other performance tests studied here, and may therefore be better suited to group-level comparisons rather than making decisions about
individual patients. This does not however, suggest that the TUG lacks reliability altogether, simply that the observed level of reliability in this study warrants caution when using the TUG for individual applications like evaluating an individual LLP user over time.

An additional consideration when interpreting test-retest ICC values is the ICC model used to estimate reliability. Model selection, which is governed both by study design and whether systematic error is included in the calculation of the reliability coefficient, determines if reliability results can be generalized beyond the participants and raters included in the study. Consistent with most reproducibility studies conducted to-date, test-retest ICCs in the present study were calculated using ICC (2,1), which includes both random and systematic error. Model 2 permits the generalization of reliability results to other raters with similar participants or patients. The reliability results of the present study can therefore be generalized to other unilateral LLP users and raters.

The SEM and MDC of the studied tests provide clinicians with indices they can use to determine precision of their patients’ scores, or determine whether changes have occurred in their patients. The SEM provides an estimate of precision in a score. The confidence interval (CI) that is derived with the SEM can be used, for example, in cross-sectional applications to assess an individual’s score relative to a cutoff score (e.g., a threshold established to indicate those at risk for falls). In contrast, MDC provides an index for use in longitudinal applications. MDC can be used to assess whether change over time in an individual patient’s performance (i.e., repeated assessments) are above and beyond changes that may be attributable to measurement error. Any change in a subject’s score, either above or below the previous score, greater than the MDC is considered a real change. For example, if a LLP user recorded an initial NBWT score of .38, completed a 4-week balance-training program, and was then re-evaluated on the NBWT with a score of .55, the difference would exceed the reported MDC of .16, indicating that a real change in balance ability had been observed.
Limitations

The present study only included established unilateral LLP users who had been using a prosthesis for greater than or equal to 16 years on average. Additional research is required to determine whether these results extend to acute LLP users (i.e., < 1 year post amputation), bilateral LLP users, or other populations. Similarly, additional research focusing on individuals with dysvascular amputations or individuals in specific age categories (e.g., the U.S. Medicare population) may be required. A larger sample size would be required to determine whether reliability and validity indices vary between sub-groups of participants (i.e., LLP users with different levels of amputation, causes of amputation, or mobility levels).

The present study only evaluated the reproducibility of four performance-based clinical tests. Given the population-specific nature of validity and reliability indices, future research is required to evaluate other psychometric properties of these tests (e.g., minimally important clinical difference), as well as those of other performance-based tests that have proven useful in other patient populations (e.g., Functional Reach Test, Fullerton Advanced Balance Scale, Five-times Sit to Stand etc).

Significant session-to-session differences were observed in the TUG. Interestingly, the direction of these differences (i.e., time decreased, performance improved) differed from those observed in a previous study, which found differences of a similar magnitude but in the opposite direction (i.e., times increased, performance decreased)\(^33\). This difference may represent the natural variation in TUG performance among LLP users. It is also possible that differences in sample characteristics could contribute. For example, there appear to be differences in age (55 versus 66 years old), level of amputation (25% versus 50% transfemoral), the length of the test-retest period, and potentially
cause of amputation between the two studies. Additional research is however required to verify these between-session differences and offer potential explanations. Regardless, these “time effects” suggest that practice effects may persist between test sessions for the TUG when it is applied with LLP users\textsuperscript{38,72,73}. While not evaluated in the present study, within-session trial-to-trial practice effects may also exist among a host of performance-based clinical tests. These practice effects may limit not only the reliability of performance-based tests, but also their validity (i.e., diagnostic accuracy) \textsuperscript{73}. Previous research on the NBWT has shown that changes to administration and scoring procedures (i.e., providing additional practice trials) can accommodate practice effects and improve psychometric performance\textsuperscript{73}. It is possible that changes to TUG administration and scoring procedures may be able to resolve these issues\textsuperscript{22}. It may therefore be prudent to examine performance-based clinical balance tests’ administration and scoring procedures to determine if they are susceptible to practice effects or other sources of systematic error\textsuperscript{38}. If they are, modifications to administration and scoring procedures may prove effective in accommodating within and between session practice effects.

Conclusion

Results of this study provide valuable evidence of inter-rater and test-retest reliability, measurement error, and minimal detectable change for performance-based clinical balance tests among established unilateral LLP users. These reliability metrics provide clinicians with actionable evidence they can use to select tests for use with individual patients, interpret the resultant test scores, and effectively use the test results to make informed clinical decisions. All studied tests demonstrated sufficient reliability to be considered appropriate for group-level comparisons (i.e., ICC ≥ .70), while the FSST, 10mWT, and NBWT were also found to be suitable for individual-level applications among established unilateral LLP users (i.e., ICC ≥ .90). SEM and MDC estimates of the studied tests provide clinicians and researchers with the information necessary to assess differences or changes in balance.
ability and/or fall risk among individual LLP users. In conjunction with published validity indices\textsuperscript{53,74}, clinicians and researchers can use the calculated values to better select these clinical tests for LLP users, and more effectively interpret the resulting scores to make informed, evidence-based clinical decisions.

Reference


Table 1. Amputation, prosthetic, and activity characteristics of the study sample

<table>
<thead>
<tr>
<th>Amputation Characteristics</th>
<th>Prosthetic Characteristics</th>
<th>Activity Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Etiology</td>
<td>Components</td>
</tr>
<tr>
<td>Transtibial (n=45)</td>
<td>Trauma (n=26)</td>
<td>Energy Storing (53)</td>
</tr>
<tr>
<td></td>
<td>Infection (n=13)</td>
<td>Single-Axis (3)</td>
</tr>
<tr>
<td></td>
<td>Dysvascular (n=11)</td>
<td>Multi-Axis (2)</td>
</tr>
<tr>
<td></td>
<td>Tumor (n=9)</td>
<td>SACH (2)</td>
</tr>
<tr>
<td></td>
<td>Other (n=1)</td>
<td>Microprocessor (13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-Microprocessor (2)</td>
</tr>
</tbody>
</table>

CI: Confidence Interval, IQR: Interquartile range, MFCL: Medicare Functional Classification Level (K-level), SACH: solid ankle cushioned heel
Table 2. Inter-rater reliability of performance-based clinical balance tests in people with lower limb loss

<table>
<thead>
<tr>
<th>Test</th>
<th>Rater A Session 1 Score Mean (SD)</th>
<th>Rater B Session 1 Score Mean (SD)</th>
<th>Score Difference Mean (SD)</th>
<th>ICC (1,1) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (score, /1.00)</td>
<td>.42 (0.21)</td>
<td>.41 (0.22)</td>
<td>0.01 (0.03)</td>
<td>.99 (.98-.99)</td>
</tr>
<tr>
<td>TUG (time, s)</td>
<td>10.4 (2.76)</td>
<td>10.4 (2.79)</td>
<td>0.05 (0.27)</td>
<td>.99 (.99-1.0)</td>
</tr>
<tr>
<td>FSST (time, s)</td>
<td>10.4 (5.27)</td>
<td>10.4 (5.21)</td>
<td>-0.03 (-0.37)</td>
<td>.99 (.99-1.0)</td>
</tr>
<tr>
<td>10MWT (time, s)</td>
<td>8.74 (1.96)</td>
<td>8.76 (1.96)</td>
<td>-0.02 (-0.30)</td>
<td>.99 (.98-.99)</td>
</tr>
</tbody>
</table>

NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement
## Table 3. Test-retest reliability of performance-based clinical balance tests in people with lower-limb loss

<table>
<thead>
<tr>
<th>Test</th>
<th>Session 1 Score Mean (SD)</th>
<th>Session 2 Score Mean (SD)</th>
<th>Score Change Mean (SD)</th>
<th>ICC (2,1) (95% CI)</th>
<th>SEM</th>
<th>MDC90 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBWT (score, /1.00)</td>
<td>.42 (.21)</td>
<td>.44 (.24)</td>
<td>.02 (.10)</td>
<td>.90 (.83-.94)</td>
<td>.07</td>
<td>.16 (.10-.20)</td>
</tr>
<tr>
<td>TUG (time, s)</td>
<td>10.4 (2.76)</td>
<td>9.91 (2.75)</td>
<td>.50 (1.30)*</td>
<td>.88 (.78-.93)</td>
<td>.96</td>
<td>2.2 (1.5-2.9)</td>
</tr>
<tr>
<td>FSST (time, s)</td>
<td>10.4 (5.27)</td>
<td>10.3 (4.83)</td>
<td>.14 (1.32)</td>
<td>.97 (.94-.98)</td>
<td>.87</td>
<td>2.0 (.79-3.5)</td>
</tr>
<tr>
<td>10MWT (time, s)</td>
<td>8.74 (1.96)</td>
<td>8.58 (2.00)</td>
<td>.16 (.54)</td>
<td>.96 (.93-.98)</td>
<td>.39</td>
<td>.91 (.40-1.4)</td>
</tr>
</tbody>
</table>

* Test score significantly different between test sessions (2-sided paired t-test) p = .004

NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement, MDC90: Minimal Detectable Change
Appendix 1

Timed Up and Go (TUG) Protocol

**Explanation to participant:** The goal of this test is to rise from the chair, walk around the cone, walk back to the chair and sit down again. You should walk at your normal, comfortable pace. I will time you while you perform the test.

**Demonstration:** Demonstrate the test one time.

**Test instructions:** Begin the test sitting with your back against the back of the chair and your arms resting on the armrests. When I say go, please stand up and walk around the cone, walk back to the chair, and sit down again. Please walk at your normal, comfortable pace.

**Practice:** Administer one practice trial. Do not time the practice trial.

**Administration:** Administer the test 2 times. Begin timing when you say go. Stop timing when the participants’ buttocks touch the chair.

**Scoring:** Select faster of the two timed trials as the TUG score.
Four Square Step Test (FSST) Protocol

Explanation to participant: The goal of this test is to step over the canes in a specific sequence as quickly as possible. I will time you while you perform this test.

Demonstration: Demonstrate the test one time. Demonstrate starting in square 1 and stepping in squares 2, 3, 4, 1, 4, 3, 2, and 1.

Test instructions: When I say go, please step in the sequent I demonstrated. Try to complete the sequence as fast as possible without touching the canes. Both feet must make contact with the floor in each square. Face forward during the entire sequence.

Practice: Administer one practice trial. Do not time the practice trial.

Administration: Administer the test 2 times. Begin timing when the first foot contacts square 2. Stop timing when the last foot contacts square 1. Repeat the trial if the participant does not complete the sequence, loses balance, or contact a cane.

Scoring: Select the faster of the two timed trials as the FSST score.
Narrowing Beam Walking Test (NBWT) Protocol

**Explanation to participant:** The goal of this test is to walk as far as possible along the beam. Speed is not being evaluated. Begin the test by standing with one foot on the wide end of the beam and the other foot on the ground to the side. You may choose which foot to put on the beam and which to put on the ground. Please cross both your arms across your chest.

**Demonstration:** Demonstrate the test one time.

**Test instructions:** When I say go, please walk along the beam as far as you can. Please walk at a comfortable speed. Remember to keep your arms crossed over your chest as you walk. Once you move your arms away from your body or step off the beam, I will ask you to stop.

**Practice:** Do not administer a practice trial.

**Administration:** Administer the test 5 times. Stop the trial when a participant: walks the length of the beam, steps off the beam, or moves their arms away from their body.

**Scoring:** Average the distances walked during trials 3 through 5, and divide by 22. This normalized distance (i.e., 0-1) is recorded as the NBWT score. The average distance is divided by 22 and not 24 (the total length of the beam), because participants begin the test with one foot on the beam and must walk past the 0.61m (2.0ft) mark to receive a score above 0.0.
10 Meter Walk Test (10MWT) Protocol

Explanation to participant: The goal of this test is to walk a short distance at your preferred comfortable walking speed. I will time you while you perform this test.

Demonstration: Demonstrate the test one time.

Test instructions: When I say ‘go,’ please walk at your normal, comfortable pace until I say stop.

Practice: Administer one practice trial.

Administration: Prepare a 14-meter walkway in a hallway or other unobstructed area. Place lines at 0, 2, 12, and 14 meters. Have the participant start in a standing position on the 0-meter line. Inform them that, on the word “go,” they are to walk at a comfortable speed until you say “stop.” Begin timing when the participant crosses the 2-meter line. Stop timing when the participant crosses the 12-meter line. Inform the participant to stop when they cross the 14-meter line. Repeat the test two times.

Scoring: Select the faster of the two timed trials as the TUG score.
<table>
<thead>
<tr>
<th>ID</th>
<th>NBWT score RaterA</th>
<th>NBWT score RaterB</th>
<th>TUG score RaterA</th>
<th>TUG score RaterB</th>
<th>FSST score RaterA</th>
<th>FSST score RaterB</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.60</td>
<td>0.60</td>
<td>8.36</td>
<td>7.94</td>
<td>4.78</td>
<td>4.89</td>
</tr>
<tr>
<td>S2</td>
<td>0.66</td>
<td>0.65</td>
<td>7.79</td>
<td>7.62</td>
<td>6.71</td>
<td>6.77</td>
</tr>
<tr>
<td>S3</td>
<td>0.83</td>
<td>0.84</td>
<td>9.20</td>
<td>9.24</td>
<td>6.94</td>
<td>7.12</td>
</tr>
<tr>
<td>S4</td>
<td>0.29</td>
<td>0.27</td>
<td>19.16</td>
<td>18.45</td>
<td>21.35</td>
<td>20.63</td>
</tr>
<tr>
<td>S5</td>
<td>0.78</td>
<td>0.78</td>
<td>7.77</td>
<td>7.63</td>
<td>6.82</td>
<td>6.96</td>
</tr>
<tr>
<td>S6</td>
<td>0.36</td>
<td>0.34</td>
<td>9.19</td>
<td>9.22</td>
<td>9.19</td>
<td>9.22</td>
</tr>
<tr>
<td>S7</td>
<td>0.29</td>
<td>0.30</td>
<td>14.02</td>
<td>14.12</td>
<td>10.94</td>
<td>11.03</td>
</tr>
<tr>
<td>S8</td>
<td>0.67</td>
<td>0.66</td>
<td>7.79</td>
<td>7.56</td>
<td>4.49</td>
<td>4.62</td>
</tr>
<tr>
<td>S9</td>
<td>0.76</td>
<td>0.76</td>
<td>7.40</td>
<td>7.77</td>
<td>6.49</td>
<td>6.36</td>
</tr>
<tr>
<td>S10</td>
<td>0.39</td>
<td>0.38</td>
<td>11.52</td>
<td>11.33</td>
<td>10.78</td>
<td>10.94</td>
</tr>
<tr>
<td>S11</td>
<td>0.27</td>
<td>0.26</td>
<td>13.42</td>
<td>13.44</td>
<td>12.96</td>
<td>12.45</td>
</tr>
<tr>
<td>S12</td>
<td>0.48</td>
<td>0.48</td>
<td>11.37</td>
<td>11.08</td>
<td>8.36</td>
<td>9.72</td>
</tr>
<tr>
<td>S13</td>
<td>0.81</td>
<td>0.81</td>
<td>8.19</td>
<td>8.02</td>
<td>6.41</td>
<td>6.26</td>
</tr>
<tr>
<td>S14</td>
<td>0.48</td>
<td>0.47</td>
<td>7.02</td>
<td>6.76</td>
<td>7.83</td>
<td>7.96</td>
</tr>
<tr>
<td>S15</td>
<td>0.30</td>
<td>0.27</td>
<td>14.82</td>
<td>14.96</td>
<td>13.90</td>
<td>14.07</td>
</tr>
<tr>
<td>S16</td>
<td>0.20</td>
<td>0.16</td>
<td>14.68</td>
<td>14.75</td>
<td>13.02</td>
<td>12.78</td>
</tr>
<tr>
<td>S17</td>
<td>0.33</td>
<td>0.33</td>
<td>9.65</td>
<td>9.64</td>
<td>9.03</td>
<td>9.22</td>
</tr>
<tr>
<td>S18</td>
<td>0.27</td>
<td>0.26</td>
<td>12.08</td>
<td>12.26</td>
<td>13.30</td>
<td>13.21</td>
</tr>
<tr>
<td>S19</td>
<td>0.34</td>
<td>0.30</td>
<td>10.63</td>
<td>10.59</td>
<td>10.06</td>
<td>10.15</td>
</tr>
<tr>
<td>S20</td>
<td>0.36</td>
<td>0.36</td>
<td>8.96</td>
<td>8.71</td>
<td>7.58</td>
<td>7.58</td>
</tr>
<tr>
<td>S21</td>
<td>0.48</td>
<td>0.48</td>
<td>10.36</td>
<td>10.35</td>
<td>9.02</td>
<td>9.02</td>
</tr>
<tr>
<td>S22</td>
<td>0.52</td>
<td>0.54</td>
<td>11.44</td>
<td>11.28</td>
<td>9.19</td>
<td>9.09</td>
</tr>
<tr>
<td>S23</td>
<td>0.41</td>
<td>0.40</td>
<td>10.67</td>
<td>10.51</td>
<td>11.69</td>
<td>11.78</td>
</tr>
<tr>
<td>S24</td>
<td>0.45</td>
<td>0.46</td>
<td>8.78</td>
<td>8.84</td>
<td>9.57</td>
<td>9.69</td>
</tr>
<tr>
<td>S25</td>
<td>0.14</td>
<td>0.13</td>
<td>20.58</td>
<td>20.72</td>
<td>32.89</td>
<td>32.94</td>
</tr>
<tr>
<td>S26</td>
<td>0.30</td>
<td>0.06</td>
<td>14.34</td>
<td>14.25</td>
<td>13.78</td>
<td>13.63</td>
</tr>
<tr>
<td>S27</td>
<td>0.25</td>
<td>0.25</td>
<td>8.65</td>
<td>8.50</td>
<td>8.39</td>
<td>8.35</td>
</tr>
<tr>
<td>S28</td>
<td>0.40</td>
<td>0.39</td>
<td>13.91</td>
<td>14.50</td>
<td>20.83</td>
<td>21.00</td>
</tr>
<tr>
<td>S29</td>
<td>0.18</td>
<td>0.18</td>
<td>9.96</td>
<td>10.01</td>
<td>17.13</td>
<td>17.21</td>
</tr>
<tr>
<td>S30</td>
<td>0.56</td>
<td>0.57</td>
<td>10.06</td>
<td>10.03</td>
<td>8.49</td>
<td>7.22</td>
</tr>
<tr>
<td>S31</td>
<td>0.69</td>
<td>0.68</td>
<td>9.71</td>
<td>9.51</td>
<td>3.33</td>
<td>3.07</td>
</tr>
<tr>
<td>S32</td>
<td>0.22</td>
<td>0.22</td>
<td>10.70</td>
<td>11.21</td>
<td>12.81</td>
<td>12.81</td>
</tr>
<tr>
<td>S33</td>
<td>0.10</td>
<td>0.08</td>
<td>11.82</td>
<td>11.51</td>
<td>14.14</td>
<td>14.52</td>
</tr>
<tr>
<td>S34</td>
<td>0.50</td>
<td>0.51</td>
<td>8.16</td>
<td>8.16</td>
<td>8.02</td>
<td>9.13</td>
</tr>
<tr>
<td>S35</td>
<td>0.39</td>
<td>0.39</td>
<td>11.35</td>
<td>11.31</td>
<td>10.52</td>
<td>10.59</td>
</tr>
<tr>
<td>S36</td>
<td>0.24</td>
<td>0.24</td>
<td>10.43</td>
<td>10.14</td>
<td>8.60</td>
<td>8.81</td>
</tr>
<tr>
<td>S37</td>
<td>0.70</td>
<td>0.70</td>
<td>7.77</td>
<td>7.57</td>
<td>7.02</td>
<td>7.02</td>
</tr>
<tr>
<td>S38</td>
<td>0.58</td>
<td>0.58</td>
<td>10.22</td>
<td>10.56</td>
<td>6.25</td>
<td>6.39</td>
</tr>
<tr>
<td>S39</td>
<td>0.51</td>
<td>0.50</td>
<td>7.69</td>
<td>7.62</td>
<td>7.00</td>
<td>7.00</td>
</tr>
<tr>
<td>S40</td>
<td>0.31</td>
<td>0.31</td>
<td>8.76</td>
<td>7.47</td>
<td>7.58</td>
<td>7.27</td>
</tr>
<tr>
<td>S41</td>
<td>0.18</td>
<td>0.17</td>
<td>10.58</td>
<td>10.40</td>
<td>11.64</td>
<td>11.70</td>
</tr>
<tr>
<td>S42</td>
<td>0.00</td>
<td>0.02</td>
<td>13.52</td>
<td>13.30</td>
<td>20.96</td>
<td>20.03</td>
</tr>
<tr>
<td>S43</td>
<td>0.80</td>
<td>0.80</td>
<td>7.35</td>
<td>7.38</td>
<td>6.39</td>
<td>6.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>S44</td>
<td>0.69</td>
<td>0.68</td>
<td>7.59</td>
<td>7.57</td>
<td>5.53</td>
<td>5.58</td>
</tr>
<tr>
<td>S45</td>
<td>0.42</td>
<td>0.42</td>
<td>8.56</td>
<td>8.68</td>
<td>8.31</td>
<td>8.40</td>
</tr>
<tr>
<td>S46</td>
<td>0.23</td>
<td>0.29</td>
<td>10.97</td>
<td>10.76</td>
<td>9.15</td>
<td>9.28</td>
</tr>
<tr>
<td>S47</td>
<td>0.38</td>
<td>0.42</td>
<td>9.43</td>
<td>9.33</td>
<td>5.65</td>
<td>6.39</td>
</tr>
<tr>
<td>S48</td>
<td>0.28</td>
<td>0.26</td>
<td>10.56</td>
<td>10.78</td>
<td>13.96</td>
<td>14.13</td>
</tr>
<tr>
<td>S49</td>
<td>0.05</td>
<td>0.05</td>
<td>11.59</td>
<td>11.50</td>
<td>12.07</td>
<td>11.81</td>
</tr>
<tr>
<td>S50</td>
<td>0.36</td>
<td>0.39</td>
<td>9.88</td>
<td>9.69</td>
<td>10.63</td>
<td>10.56</td>
</tr>
<tr>
<td>S51</td>
<td>0.61</td>
<td>0.61</td>
<td>7.34</td>
<td>7.20</td>
<td>4.15</td>
<td>4.14</td>
</tr>
<tr>
<td>S52</td>
<td>0.19</td>
<td>0.16</td>
<td>8.00</td>
<td>8.04</td>
<td>8.46</td>
<td>8.63</td>
</tr>
<tr>
<td>S53</td>
<td>0.55</td>
<td>0.57</td>
<td>8.01</td>
<td>8.03</td>
<td>5.71</td>
<td>5.78</td>
</tr>
<tr>
<td>S54</td>
<td>0.21</td>
<td>0.19</td>
<td>12.31</td>
<td>12.46</td>
<td>11.45</td>
<td>11.45</td>
</tr>
<tr>
<td>S55</td>
<td>0.39</td>
<td>0.43</td>
<td>9.76</td>
<td>10.08</td>
<td>8.39</td>
<td>8.39</td>
</tr>
<tr>
<td>S56</td>
<td>0.58</td>
<td>0.58</td>
<td>9.70</td>
<td>9.68</td>
<td>8.53</td>
<td>8.64</td>
</tr>
<tr>
<td>S57</td>
<td>0.48</td>
<td>0.48</td>
<td>7.97</td>
<td>8.07</td>
<td>8.97</td>
<td>8.95</td>
</tr>
<tr>
<td>S58</td>
<td>0.64</td>
<td>0.65</td>
<td>10.06</td>
<td>10.01</td>
<td>11.57</td>
<td>11.82</td>
</tr>
<tr>
<td>S59</td>
<td>0.16</td>
<td>0.16</td>
<td>14.84</td>
<td>14.83</td>
<td>25.77</td>
<td>25.93</td>
</tr>
<tr>
<td>S60</td>
<td>0.23</td>
<td>0.22</td>
<td>8.34</td>
<td>8.46</td>
<td>8.70</td>
<td>8.81</td>
</tr>
<tr>
<td>10MWT score</td>
<td>NBWT score</td>
<td>TUG score</td>
<td>FSST score</td>
<td>10MWT score</td>
<td>SEX</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>-----------</td>
<td>------------</td>
<td>------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>RaterA</td>
<td>RaterB</td>
<td>RaterA</td>
<td>RaterA</td>
<td>RaterA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td>7.13</td>
<td>0.78</td>
<td>8.60</td>
<td>5.09</td>
<td>7.39</td>
<td>male</td>
</tr>
<tr>
<td>7.21</td>
<td>7.10</td>
<td>0.84</td>
<td>7.32</td>
<td>6.00</td>
<td>6.96</td>
<td>female</td>
</tr>
<tr>
<td>7.00</td>
<td>7.26</td>
<td>0.75</td>
<td>8.19</td>
<td>5.71</td>
<td>6.44</td>
<td>female</td>
</tr>
<tr>
<td>9.89</td>
<td>10.06</td>
<td>0.16</td>
<td>12.59</td>
<td>18.90</td>
<td>9.46</td>
<td>male</td>
</tr>
<tr>
<td>5.62</td>
<td>5.52</td>
<td>0.85</td>
<td>7.51</td>
<td>6.39</td>
<td>5.52</td>
<td>male</td>
</tr>
<tr>
<td>8.50</td>
<td>8.49</td>
<td>0.41</td>
<td>9.01</td>
<td>9.31</td>
<td>8.55</td>
<td>female</td>
</tr>
<tr>
<td>11.84</td>
<td>12.02</td>
<td>0.30</td>
<td>14.09</td>
<td>10.87</td>
<td>12.23</td>
<td>female</td>
</tr>
<tr>
<td>7.10</td>
<td>7.13</td>
<td>0.66</td>
<td>7.80</td>
<td>4.62</td>
<td>6.98</td>
<td>female</td>
</tr>
<tr>
<td>6.03</td>
<td>5.96</td>
<td>0.70</td>
<td>7.97</td>
<td>6.16</td>
<td>5.82</td>
<td>female</td>
</tr>
<tr>
<td>9.96</td>
<td>9.90</td>
<td>0.43</td>
<td>11.15</td>
<td>11.77</td>
<td>9.59</td>
<td>female</td>
</tr>
<tr>
<td>13.05</td>
<td>13.25</td>
<td>0.27</td>
<td>12.90</td>
<td>15.85</td>
<td>11.70</td>
<td>male</td>
</tr>
<tr>
<td>9.43</td>
<td>9.72</td>
<td>0.43</td>
<td>10.56</td>
<td>7.63</td>
<td>9.52</td>
<td>male</td>
</tr>
<tr>
<td>7.45</td>
<td>7.52</td>
<td>0.73</td>
<td>7.22</td>
<td>5.78</td>
<td>7.70</td>
<td>male</td>
</tr>
<tr>
<td>7.18</td>
<td>7.62</td>
<td>0.48</td>
<td>7.25</td>
<td>7.77</td>
<td>7.18</td>
<td>male</td>
</tr>
<tr>
<td>10.89</td>
<td>10.77</td>
<td>0.30</td>
<td>15.12</td>
<td>14.01</td>
<td>10.73</td>
<td>male</td>
</tr>
<tr>
<td>9.78</td>
<td>9.84</td>
<td>0.24</td>
<td>15.11</td>
<td>13.02</td>
<td>8.66</td>
<td>male</td>
</tr>
<tr>
<td>6.35</td>
<td>7.28</td>
<td>0.33</td>
<td>9.77</td>
<td>9.11</td>
<td>7.03</td>
<td>male</td>
</tr>
<tr>
<td>10.03</td>
<td>9.94</td>
<td>0.28</td>
<td>12.09</td>
<td>13.66</td>
<td>10.01</td>
<td>male</td>
</tr>
<tr>
<td>8.62</td>
<td>8.76</td>
<td>0.33</td>
<td>11.07</td>
<td>9.92</td>
<td>8.55</td>
<td>male</td>
</tr>
<tr>
<td>6.76</td>
<td>6.90</td>
<td>0.42</td>
<td>9.00</td>
<td>7.32</td>
<td>6.53</td>
<td>female</td>
</tr>
<tr>
<td>9.77</td>
<td>9.83</td>
<td>0.46</td>
<td>10.13</td>
<td>9.49</td>
<td>9.62</td>
<td>male</td>
</tr>
<tr>
<td>11.88</td>
<td>11.64</td>
<td>0.51</td>
<td>11.87</td>
<td>9.13</td>
<td>11.54</td>
<td>male</td>
</tr>
<tr>
<td>7.94</td>
<td>8.01</td>
<td>0.49</td>
<td>10.63</td>
<td>11.93</td>
<td>7.55</td>
<td>male</td>
</tr>
<tr>
<td>8.37</td>
<td>8.34</td>
<td>0.42</td>
<td>8.31</td>
<td>9.43</td>
<td>8.81</td>
<td>male</td>
</tr>
<tr>
<td>17.26</td>
<td>17.06</td>
<td>0.17</td>
<td>20.11</td>
<td>31.77</td>
<td>18.22</td>
<td>male</td>
</tr>
<tr>
<td>10.15</td>
<td>10.27</td>
<td>0.14</td>
<td>15.33</td>
<td>13.83</td>
<td>10.03</td>
<td>male</td>
</tr>
<tr>
<td>8.01</td>
<td>7.96</td>
<td>0.33</td>
<td>8.72</td>
<td>8.22</td>
<td>8.13</td>
<td>male</td>
</tr>
<tr>
<td>10.45</td>
<td>11.06</td>
<td>0.45</td>
<td>13.80</td>
<td>20.19</td>
<td>10.41</td>
<td>male</td>
</tr>
<tr>
<td>7.90</td>
<td>8.15</td>
<td>0.17</td>
<td>10.02</td>
<td>16.32</td>
<td>7.78</td>
<td>male</td>
</tr>
<tr>
<td>8.63</td>
<td>8.38</td>
<td>0.52</td>
<td>10.21</td>
<td>7.13</td>
<td>8.22</td>
<td>male</td>
</tr>
<tr>
<td>8.20</td>
<td>8.21</td>
<td>0.87</td>
<td>3.56</td>
<td>8.96</td>
<td>8.53</td>
<td>male</td>
</tr>
<tr>
<td>9.33</td>
<td>9.33</td>
<td>0.14</td>
<td>11.22</td>
<td>11.76</td>
<td>9.19</td>
<td>male</td>
</tr>
<tr>
<td>9.89</td>
<td>9.82</td>
<td>0.27</td>
<td>9.22</td>
<td>14.69</td>
<td>11.52</td>
<td>male</td>
</tr>
<tr>
<td>9.19</td>
<td>7.63</td>
<td>0.58</td>
<td>7.09</td>
<td>8.10</td>
<td>8.62</td>
<td>female</td>
</tr>
<tr>
<td>9.72</td>
<td>9.69</td>
<td>0.30</td>
<td>9.62</td>
<td>10.28</td>
<td>9.93</td>
<td>male</td>
</tr>
<tr>
<td>9.64</td>
<td>9.53</td>
<td>0.28</td>
<td>11.22</td>
<td>9.26</td>
<td>9.25</td>
<td>male</td>
</tr>
<tr>
<td>7.96</td>
<td>7.65</td>
<td>0.72</td>
<td>7.88</td>
<td>7.50</td>
<td>6.13</td>
<td>male</td>
</tr>
<tr>
<td>10.50</td>
<td>11.03</td>
<td>0.74</td>
<td>8.87</td>
<td>7.31</td>
<td>9.62</td>
<td>male</td>
</tr>
<tr>
<td>6.72</td>
<td>6.62</td>
<td>0.55</td>
<td>7.46</td>
<td>8.08</td>
<td>7.08</td>
<td>female</td>
</tr>
<tr>
<td>7.27</td>
<td>7.19</td>
<td>0.55</td>
<td>8.87</td>
<td>7.41</td>
<td>6.72</td>
<td>female</td>
</tr>
<tr>
<td>8.08</td>
<td>8.03</td>
<td>0.09</td>
<td>10.00</td>
<td>11.09</td>
<td>7.83</td>
<td>female</td>
</tr>
<tr>
<td>11.01</td>
<td>11.22</td>
<td>0.02</td>
<td>13.33</td>
<td>17.12</td>
<td>9.96</td>
<td>male</td>
</tr>
<tr>
<td>6.41</td>
<td>6.50</td>
<td>0.97</td>
<td>7.46</td>
<td>5.81</td>
<td>6.38</td>
<td>female</td>
</tr>
<tr>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>8.15</td>
<td>8.13</td>
<td>0.79</td>
<td>7.08</td>
<td>6.13</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td>8.60</td>
<td>8.71</td>
<td>0.39</td>
<td>8.08</td>
<td>7.75</td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>8.44</td>
<td>8.40</td>
<td>0.23</td>
<td>9.76</td>
<td>9.81</td>
<td>8.16</td>
<td></td>
</tr>
<tr>
<td>8.63</td>
<td>8.58</td>
<td>0.46</td>
<td>6.94</td>
<td>5.95</td>
<td>7.56</td>
<td></td>
</tr>
<tr>
<td>9.13</td>
<td>8.84</td>
<td>0.18</td>
<td>10.66</td>
<td>15.59</td>
<td>9.02</td>
<td></td>
</tr>
<tr>
<td>8.25</td>
<td>8.06</td>
<td>0.14</td>
<td>11.56</td>
<td>12.39</td>
<td>8.02</td>
<td></td>
</tr>
<tr>
<td>9.26</td>
<td>9.20</td>
<td>0.43</td>
<td>8.77</td>
<td>10.94</td>
<td>8.44</td>
<td></td>
</tr>
<tr>
<td>6.83</td>
<td>6.84</td>
<td>0.59</td>
<td>7.25</td>
<td>3.97</td>
<td>7.06</td>
<td></td>
</tr>
<tr>
<td>7.15</td>
<td>6.96</td>
<td>0.37</td>
<td>7.78</td>
<td>7.70</td>
<td>6.76</td>
<td></td>
</tr>
<tr>
<td>7.56</td>
<td>7.51</td>
<td>0.84</td>
<td>8.34</td>
<td>5.19</td>
<td>7.65</td>
<td></td>
</tr>
<tr>
<td>7.32</td>
<td>7.47</td>
<td>0.18</td>
<td>11.13</td>
<td>9.27</td>
<td>7.51</td>
<td></td>
</tr>
<tr>
<td>7.81</td>
<td>7.84</td>
<td>0.41</td>
<td>8.84</td>
<td>8.34</td>
<td>7.64</td>
<td></td>
</tr>
<tr>
<td>7.81</td>
<td>7.63</td>
<td>0.37</td>
<td>9.33</td>
<td>8.26</td>
<td>7.88</td>
<td></td>
</tr>
<tr>
<td>6.60</td>
<td>6.56</td>
<td>0.36</td>
<td>8.59</td>
<td>8.44</td>
<td>7.08</td>
<td></td>
</tr>
<tr>
<td>8.78</td>
<td>8.75</td>
<td>0.74</td>
<td>9.69</td>
<td>12.27</td>
<td>8.28</td>
<td></td>
</tr>
<tr>
<td>12.20</td>
<td>12.40</td>
<td>0.21</td>
<td>14.45</td>
<td>21.78</td>
<td>11.96</td>
<td></td>
</tr>
<tr>
<td>8.08</td>
<td>8.59</td>
<td>0.26</td>
<td>7.35</td>
<td>7.33</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>LEVEL</td>
<td>ETIOLOGY</td>
<td>K-LEVEL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>TF</td>
<td>infection</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>TF</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>TT</td>
<td>cancer</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>TT</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>TT</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>TT</td>
<td>dysvascular</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>TT</td>
<td>dysvascular</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>TF</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>TT</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>TT</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>TF</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>TF</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>TF</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>TF</td>
<td>infection</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>TF</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>TT</td>
<td>infection</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>TT</td>
<td>dysvascular</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>TT</td>
<td>trauma</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>TT</td>
<td>dysvascular</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>TF</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>TT</td>
<td>infection</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>TT</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>TT</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>TF</td>
<td>infection</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>TT</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>TT</td>
<td>cancer</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>TT</td>
<td>trauma</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>TT</td>
<td>other</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>TF</td>
<td>trauma</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>TF</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>TF</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>TT</td>
<td>infection</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>TF</td>
<td>cancer</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>TT</td>
<td>trauma</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>TT</td>
<td>infection</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>TT</td>
<td>trauma</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>TT</td>
<td>trauma</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>TF</td>
<td>cancer</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>TT</td>
<td>dysvascular</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>