

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.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE Oct 2019		2. REPORT TYPE Annual		3. DATES COVERED 29 Sep 2018 – 28 Sep 2019	
4. TITLE AND SUBTITLE Can a novel beam-walking test improve fall risk assessment in Service members, Veterans, and civilians who use lower-limb prostheses?				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-17-1-0547	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) ADr. Andrew Sawers E-Mail: asawers@uic.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Illinois 809 S MARSHFIELD RM 520 CHICAGO IL 60612-4305				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. 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.					
15. SUBJECT TERMS amputee; balance; falls; recruitment; enrollment					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT U Unclassified	b. ABSTRACT U Unclassified	c. THIS PAGE U Unclassified			19b. TELEPHONE NUMBER ² (include area code)

Table of Contents

	<u>Page</u>
1. Introduction.....	4
2. Keywords.....	4
3. Accomplishments.....	4-9
4. Impact.....	9
5. Changes/Problems.....	9
6. Products, Inventions, Patent Applications, and/or Licenses.....	9-10
7. Participants & Other Collaborating Organizations.....	11
8. Special Reporting Requirements.....	12
9. Appendices.....	12

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

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 a 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 two faller classifications

Test	AUC (95% CI)	Cutoff Score	≥ 1 falls (any falls)		LR+ (95% CI)	LR- (95% CI)
			Sensitivity (95% CI)	Specificity (95% CI)		
NBWT (/1.00)	.81 (.62-.91)	≤ 0.43	73% (53%-90%)	76% (54%-96%)	3.0 (1.5-6.9)	.36 (.19-.77)
TUG (sec)	.71 (.54-.88)	≥ 8.17	83% (68%-98%)	68% (46%-92%)	2.6 (1.3-5.6)	.24 (.13-.56)
FSST (sec)	.70 (.53-.86)	≥ 8.49	74% (58%-92%)	68% (46%-92%)	2.4 (1.1-5.2)	.36 (.17-.78)
BBS (/56)	.66 (.47-.83)	≤ 50.5	67% (48%-86%)	62% (39%-86%)	1.8 (.89-3.6)	.53 (.27-1.1)
ABC (/100)	.65 (.47-.82)	≤ 80.2	50% (30%-70%)	74% (54%-96%)	1.9 (.78-5.1)	.67 (.41-1.1)
Model 1	.84 (.71-.96)	N/A	80% (64%-96%)	73% (51%-96%)	3.0 (1.3-7.1)	.27 (.12-.63)
≥ 2 falls (multiple falls)						
Test	AUC (95% CI)	Cutoff Score	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
NBWT (/1.00)	.89 (.78-1.0)	≤ 0.43	88% (73%-100%)	79% (65%-92%)	4.2 (1.7-6.9)	.16 (.042-.60)
TUG (sec)	.88 (.78-.98)	≥ 9.25	82% (64%-100%)	83% (67%-98%)	4.7 (1.9-11.9)	.21 (.075-.61)
FSST (sec)	.88 (.76-.99)	≥ 8.71	94% (83%-100%)	74% (60%-92%)	3.6 (1.8-7.3)	.08 (.012-.54)
BBS (/56)	.81 (.66-.92)	≤ 50.5	88% (73%-100%)	70% (51%-88%)	2.9 (1.5-5.5)	.17 (.045-.64)
ABC (/100)	.71 (.54-.87)	≤ 80.2	65% (42%-87%)	78% (61%-95%)	3.0 (1.3-7.0)	.45 (.23-.89)
Model 2	.90 (.79-1.0)	N/A	76% (56%-97%)	83% (67%-98%)	4.4 (1.7-11.3)	.28 (.12-.68)

ABC: Activities-specific Balance Confidence Scale; AUC: Area Under the Curve; BBS: Berg Balance Scale; CI: Confidence Interval; FSST: Four Square Step test; TUG: Timed Up and Go.

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 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: i) select balance tests that are suitable for

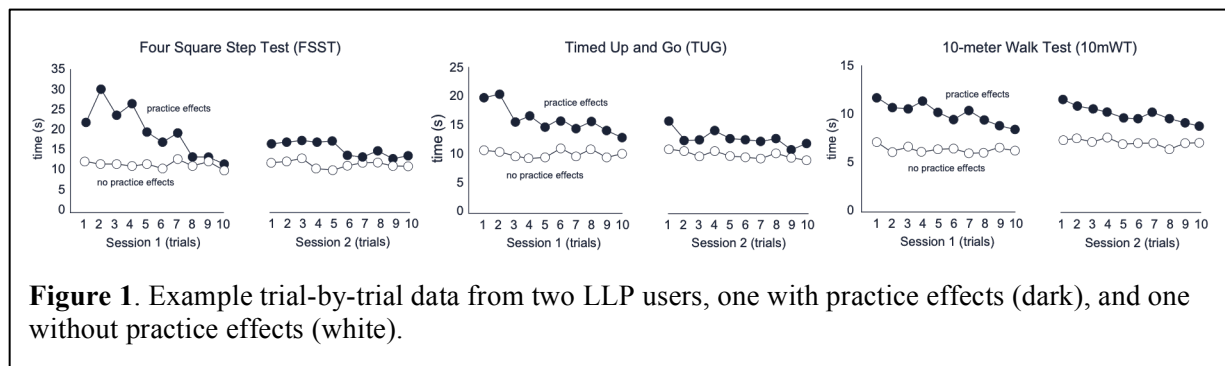
making group-level comparisons or individual-level clinical decisions (i.e., test-retest intra-class correlation coefficients), ii) interpret the precision of individual test scores on performance-based clinical balance tests (i.e., standard error of measurement), and iii) evaluate change over time in 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

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

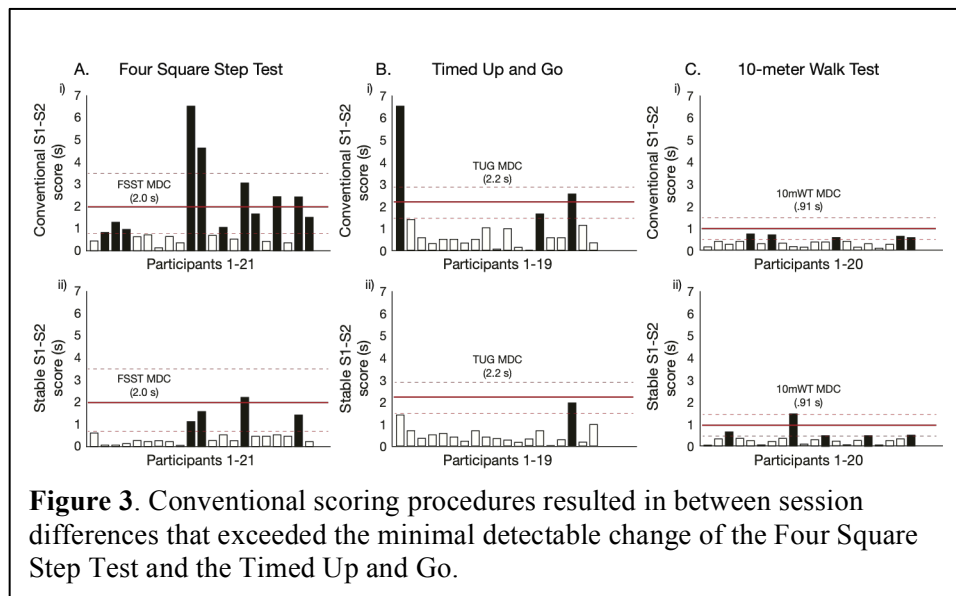
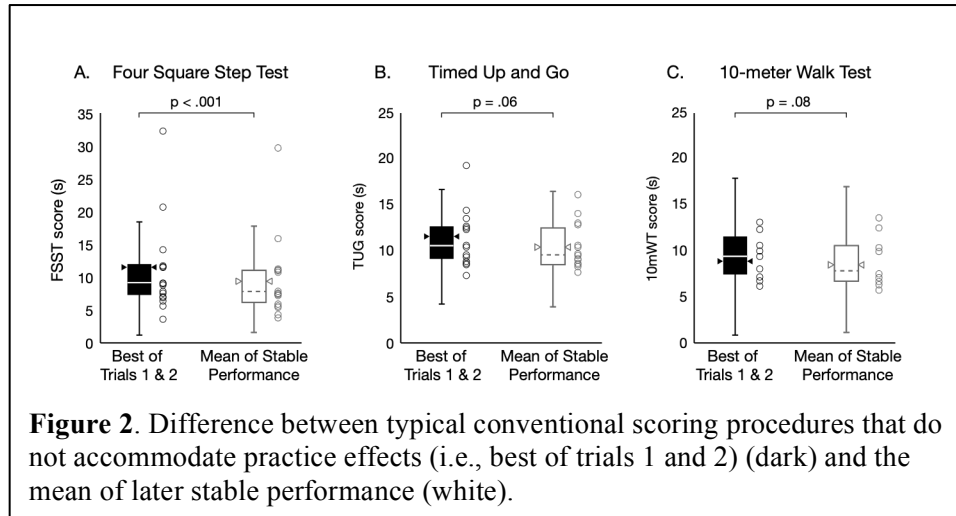
* Test score significantly different between test sessions (2-sided paired t-test) p=.004. NBWT: Narrowing Beam Walking Test, TUG: Timed Up and Go, FSST: Four Square Step Test, 10MWT: 10 Meter Walk Test, s: seconds, SD: Standard Deviation, CI: Confidence Interval, SEM = Standard Error of Measurement, MDC₉₀: Minimal Detectable Change

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).



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)

Sawers A, Hafner BJ. (2019) Conventional administration and scoring procedures suppress the diagnostic accuracy of a performance-based test designed to assess balance ability in lower limb prosthesis users. *Prosthet Orthot Int.* 43(4):402-408.

Sawers A, Hafner BJ. (2019) Using clinical balance tests to assess fall risk in established lower limb prosthesis users: cutoff scores and related validity indices. *PM & R.* (in press).

Sawers A, Kim J, Balkman GS, Hafner BJ. (2019) Inter-rater and test-retest reliability of performance-based clinical tests administered to unilateral lower limb prosthesis users. *Physical Therapy* 2019 (accepted).

Sawers A, Kim J, Balkman GS, Hafner BJ. (submitted) Characterizing practice effects in performance-based clinical tests administered to lower limb prosthesis users.

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, Kim J, Balkman GS, Hafner BJ. Orthotic and Prosthetic Innovative Technologies Conference. Evidence-based selection, use, and interpretation of performance-based clinical balance tests to improve fall risk assessment in lower limb prosthesis users. Ann Arbor MI, 2019 (*invited speaker*).

Sawers A, Kim J, Balkman GS, Hafner BJ. Scheck and Siress Prosthetics and Orthotics Education Fair. Reducing barriers to clinical selection, use, and interpretation of performance-based clinical balance tests. Chicago IL, 2019 (*invited speaker*).

Sawers A, Kim J, Major MJ, Hafner BJ. 45th American Academy of Orthotists and Prosthetists Annual Meeting and Scientific Symposium. Falls by lower limb prosthesis users: Refocusing through an epidemiological lens. Orlando, FL, 2019 (*invited symposium*).

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*).

Sawers, A., Kim, J., Hafner, B. Exploring practice effects in performance-based clinical tests administered to lower limb prosthesis users: Potential clinical significance. 8th Annual movement and Rehabilitation Science Training Day, Chicago, IL, 2019 (*Poster*).

Sawers A, Kim J, Balkman G, Hafner BJ. Deriving Reproducibility Indices To Facilitate Evidence-based Selection And Interpretation Of Performance-based Balance Tests For Lower Limb Prosthesis Users. Military Health System Research Symposium. Kissimmee, FL, 2019 (*Podium*).

Sawers A, Kim J, Hafner BJ. Administration and scoring procedures for performance-based clinical balance tests do not accommodate practice effects among lower limb prosthesis users. International Society of Posture and Gait Research World Congress. Edinburgh, Scotland, 2019 (*Poster*).

Kim J, Major MJ, Hafner BJ, Sawers A. The frequency and circumstances of falls reported by unilateral lower limb prosthesis users. International Society of Posture and Gait Research World Congress Edinburgh, Scotland, 2019 (*Poster*).

Sawers A, Kim J, Hafner BJ. Practice-effects in performance-based clinical balance tests limit their application among lower limb prosthesis users. American Congress of Rehabilitation Medicine 96th Annual Conference. Chicago, IL 2019 (*Podium*).

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

Name:	Andrew Sawers, PhD
Project Role:	Principal Investigator (UIC)
Researcher Identifier:	
Nearest person month worked:	1
Contribution to Project:	Dr. Sawers has been responsible for overseeing all aspects of the project (IRB, managing recruitment, enrollment, and data collection, manuscript preparation)
Funding Support:	N/A

Name:	Brian Hafner, PhD
Project Role:	Co-Investigator (UW)
Researcher Identifier:	
Nearest person month worked:	1
Contribution to Project:	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).
Funding Support:	N/A

Name:	Janis Kim, MPT
Project Role:	Graduate Student (UIC)
Researcher Identifier:	
Nearest person month worked:	1
Contribution to Project:	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.
Funding Support:	N/A

Name:	Geoff Balkman
Project Role:	Graduate Student (UW)
Researcher Identifier:	
Nearest person month worked:	<1
Contribution to Project:	Mr. Balkman has been responsible for data collection and data entry at the University of Washington study site.
Funding Support:	N/A

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

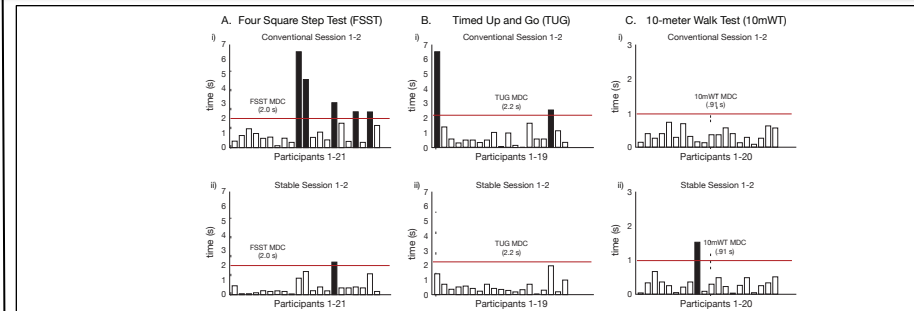


Study Aim(s)

- 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).



Accomplishment: Conventional scoring procedures resulted in between session differences (i.e., test-retest) that exceeded the minimal detectable change (MDC_{90}) 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.

Timeline and Cost

Activities	CY	17	18	19
Human subjects approval & train sites		■		
Participant recruitment			■	■
Conduct data collection procedure		■	■	■
Analyze data and disseminate results			■	■
Estimated Budget (\$K)	\$300	\$75	\$150	\$75

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)

Updated: (01/07/2010)

Conventional administration and scoring procedures suppress the diagnostic accuracy of a performance-based test designed to assess balance ability in lower limb prosthesis users

Prosthetics and Orthotics International

2019, Vol. 43(4) 402–408

© The International Society for

Prosthetics and Orthotics 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0309364619848274

journals.sagepub.com/home/poiAndrew 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

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

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

Corresponding author:

Andrew Sawers, Department of Kinesiology, University of Illinois at Chicago, 1919 West Taylor Street, Rm. 646, Chicago, IL 60612, USA.
Email: asawers@uic.edu

Background

Over 50% of lower limb prosthesis (LLP) users report falling at least once a year,^{1–3} placing them at a high risk for adverse health outcomes that include injury,^{2,4} reduced mobility,⁵ and diminished quality of life.^{6,7} 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.⁸ Existing performance-based clinical balance tests do not generally pose sufficient challenge to identify differences in balance control that underlie fall risk,⁹ reducing their ability to discriminate between LLP users with and without a history of falls.⁸

A standardized performance-based clinical balance test termed the Narrowing Beam Walking Test (NBWT) was developed to address this limitation among LLP users.^{10,11} 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.^{8,11} 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.⁸ 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.^{8,12}

Beyond increasing the challenge to balance control, development of the NBWT included the study practice effects associated with beam-walking performance by LLP users.^{9,11} 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.^{13,14} 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,^{15–20} few clinical test developers explicitly state whether their recommended administration and scoring procedures address possible practice effects,²¹ 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.¹¹ 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 tests¹¹ (e.g. mean or best of trials 1–3).^{22–25} It remains unknown therefore whether the superior diagnostic accuracy of the NBWT among LLP users⁸ 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 3–5, 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)^{15,26,27} 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 study⁷ 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 study¹¹ 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

Table 1. Sociodemographic and prosthetic-related characteristics.

	Age (years)	Height (cm)	Weight (kg)	Years Since Limb Loss	Sex	Amputation Level	Amputation Etiology	MFCL
Mean (SD)	48.7 (14.6)	173 (9.10)	78.5 (14.1)	14.3 (12.6)	M (21)	TT (25)	Trauma (25)	K1 (2)
Range	24–70	155–189	47–107	2–57	F (19)	TF (15)	Dysvascular (7) Cancer (5) Infection (3)	K2 (15) K3 (17) K4 (6)

MFCL: Medicare Functional Classification Level (K-level); M: male; F: female; TT: transtibial; TF: transfemoral.

control, and potential balance, may be fundamentally different than someone who loses 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 ($\alpha < .0167$) 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 $\pm$ 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, $\chi^2(2)=6.508$,

Table 2. Participant characteristics potentially contributing to practice effects.

	Age (years)	NBWT baseline first trial	MFCL*		Amputation level*		Amputation etiology*		Fall history*	
	Mean (SD)	Mean (SD)	K1&2	K3&4	TT	TF	DV	NDV	Fall	No Fall
Practice effects (n = 17)	49.35 (12.52)	0.405 (0.232)	7	10	13	4	1	16	8	9
No practice effects (n = 23)	48.26 (16.17)	0.290 (0.209)	10	13	12	11	7	16	16	7
p-value	0.818	0.109	1.00		0.187		0.107		0.199	

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.

$p=0.039$. Using the Greenhouse-Geisser correction, NBWT scores were found to be significantly different between scoring methods, $F(1.73, 67.39)=16.87$, $p<0.001$. After a Bonferroni correction, the mean of trials 3–5 was significantly lower than the best of trials 1–3, $p=0.041$, and significantly greater than the mean of trials 1–3, $p=0.017$. The best of trials 1–3 was also significantly greater than the mean of trials 1–3, $p<0.001$.

The area under the ROC curve (area, 95% CI) was larger for the mean of trials 3–5 (0.81, 0.62–0.91) than for the best of trials 1–3 (0.67, 0.50–0.84) or the mean of trials 1–3 (0.69, 0.52–0.85). The area under the ROC curve was significantly greater than 0.50 for the mean of trials 3–5 ($p=0.006$) and the mean of trials 1–3 ($p=0.041$), but not for the best of trials 1–3 ($p=0.66$). Only the mean of trials 3–5 had an area under the ROC curve that exceeded 0.80 (Supplemental Figure 1). No significant differences ($p<0.05$) in factors previously reported to influence practice effects were identified between LLP users who did and did not exhibit practice effects on the NBWT (Table 2). Similarly, LLP users who exhibited practice effects had a greater NBWT score (i.e. mean of trials 3–5: 0.458) than those who did not (i.e. 0.361); however, this difference was not statistically significant ($p=0.118$).

Discussion

This study was conducted to test the hypothesis that scoring the NBWT using conventional methods (i.e. mean or best of trials 1–3) would adversely affect its diagnostic accuracy. Results show that the diagnostic accuracy of the NBWT was reduced significantly 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 perfor-

mance-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 users⁸ 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 accuracy³⁶ 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.³⁶

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 practice 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

- Kim J, Hafner BJ, Major MJ, et al. Frequency and circumstances of falls reported by ambulatory unilateral lower limb prosthesis users: a secondary analysis. *PM R* 2019; 11: 344–353.
- Kulkarni J, Wright S, Toole C, et al. Falls in patients with lower limb amputations: prevalence and contributing factors. *Physiotherapy* 1996; 82: 130–136.
- Miller WC, Speechley M and Deathe B. The prevalence and risk factors of falling and fear of falling among lower extremity amputees. *Arch Phys Med Rehabil* 2001; 82(8): 1031–1037.
- Mundell B, MaraditKremers H, Visscher S, et al. Direct medical costs of accidental falls for adults with transfemoral amputations. *Prosthet Orthot Int* 2017; 41(6): 564–570.
- van Velzen JM, van Bennekom CA, Polonski W, et al. Physical capacity and walking ability after lower limb amputation: a systematic review. *Clin Rehabil* 2006; 20(11): 999–1016.
- Pezzin LE, Dillingham TR and MacKenzie EJ. Rehabilitation and the long-term outcomes of persons with trauma-related amputations. *Arch Phys Med Rehabil* 2000; 81(3): 292–300.
- Barnett CT, Vanicek N and Polman RCJ. Temporal adaptations in generic and population-specific quality of life and falls efficacy in men with recent lower-limb amputations. *J Rehabil Res Dev* 2013; 50(3): 437–48.
- Sawers A and Hafner B. Validation of the narrowing beam walking test in lower limb prosthesis users. *Arch Phys Med Rehabil* 2018; 99(8): 1491–1498.e1.
- Sawers A and Ting LH. Beam walking can detect differences in walking balance proficiency across a range of sensorimotor abilities. *Gait Posture* 2015; 41(2): 619–623.
- Sawers A and Hafner BJ. A study to assess whether fixed-width beam walking provides sufficient challenge to assess balance ability across lower limb prosthesis users. *Clin Rehabil* 2018; 32(4): 483–492.
- Sawers A and Hafner BJ. Narrowing beam walking is a clinically feasible approach for assessing balance ability in lower-limb prosthesis users. *J Rehabil Med* 2018; 50(5): 457–464.
- Sawers A, Allen JL and Ting LH. Long-term training modifies the modular structure and organization of walking balance control. *J Neurophysiol* 2015; 114(6): 3359–3373.
- Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000; 30(1): 1–15.
- Bartels C, Wegrzyn M, Wiedl A, et al. Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neurosci* 2010; 11: 118–112.
- Guyatt GH, Pugsley SO, Sullivan MJ, et al. Effect of encouragement on walking test performance. *Thorax* 1984; 39(11): 818–822.
- Halsne EG, McDonald CL, Morgan SJ, et al. Assessment of low- and high-level task performance in people with transtibial amputation using crossover and energy-storing prosthetic feet: a pilot study. *Prosthet Orthot Int* 2018; 42: 583–591.
- Kristensen MT, Ekdahl C, Kehlet H, et al. How many trials are needed to achieve performance stability of the timed up & Go test in patients with hip fracture? *Arch Phys Med Rehabil* 2010; 91: 885–889.
- Bloch ML, Jönsson LR and Kristensen MT. Introducing a third timed up & go test trial improves performances of hospitalized and community-dwelling older individuals. *J Geriatr Phys Ther* 2007; 40: 121–126.
- Lin SJ and Bose NH. Six-minute walk test in persons with transtibial amputation. *Arch Phys Med Rehabil* 2008; 89(12): 2354–2359.
- Kristensen MT, Nielsen AØ, Topp UM, et al. Number of test trials needed for performance and interrater reliability of the one leg stand test in patients with a major non-traumatic lower limb amputation. *Gait Posture* 2014; 39: 424–429.
- Ashendorf L and Horwitz JE. Abbreviating the finger tapping test. *Arch Clin Neuropsychol* 2015; 30(2): 99–104.
- Podsiadlo D and Richardson S. The timed “up & go”: a test of basic functional mobility for frail elderly persons. *JAGS* 1991; 39(2): 142–148.
- Dite W, Connor HJ and Curtis HC. Clinical identification of multiple fall risk early after unilateral transtibial amputation. *Arch Phys Med Rehabil* 2007; 88(1): 109–114.
- Deathe AB and Miller WE. The L test of functional mobility: measurement properties of a modified version of the timed “up & go” test designed for people with lower-limb amputations. *Phys Ther* 2005; 85(7): 625–635.
- Wolfson L, Whipple R, Derby C, et al. Balance and strength training in older adults: intervention gains and Tai Chi maintenance. *J Am Geriatr Soc* 1996; 44(5): 498–506.
- Adsett J, Mullins R, Hwang R, et al. Repeat six-minute walk tests in patients with chronic heart failure: are they clinically necessary. *Eur J Cardiovasc Prev Rehabil* 2011; 18(4): 601–606.

27. Liu J, Drutz C, Kumar R, et al. Use of the six-minute walk test post-stroke: is there a practice effect? *Arch Phys Med Rehabil* 2008; 89: 1686–1692.
28. Askham J, Glucksman E, Owens P, et al. *Home and leisure accident research: a review of research on falls among elderly people*. London: Age Concern Institute of Gerontology, King's College London, and Department of Trade and Industry, 1990.
29. Hauer K, Lamb SE, Jorstad EC, et al. Systematic review of definitions and methods of measuring falls in randomized controlled fall prevention trials. *Age and Ageing* 2006; 35(1): 5–10.
30. Lamb SE, Jorstad-Stein EC, Hauer K, et al. Development of a common outcome data set for fall injury prevention trials: the prevention of falls network Europe consensus. *J Am Geriatr Soc* 2005; 53(9): 1618–1622.
31. Wong C, Chihuri S and Li G. Risk of fall-related injury in people with lower limb amputations: a prospective cohort study. *J Rehabil Med* 2016; 48(1): 80–85.
32. Topper AK, Maki BE and Holliday PJ. Are activity-based assessments of balance and gait in the elderly predictive of risk of falling and/or type of fall? *JAGS* 1993; 141(5): 479–487.
33. Berg WP, Alessio HM, Mills EM, et al. Circumstances and consequences of falls in independent community-dwelling older adults. *Age Ageing* 1997; 26(4): 261–268.
34. Gallistel CR, Fairhurst S and Balsam P. The learning curve: implications of a quantitative analysis. *Proc Natl Acad Sci U S A* 2004; 101(36): 13124–13131.
35. Hanley JA and McNeil BJ. A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology* 1983; 148(3): 839–843.
36. Haynes RB. *Clinical epidemiology: how we do clinical practice research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2006.
37. Bandinelli S, Pozzi M, Lauretani F, et al. Adding challenge to performance-based tests of walking. *Am J Phys Med Rehabil* 2006; 85(12): 986–991.
38. McCaffrey RJ and Westervelt HJ. Issues associated with repeated neuropsychological assessments. *Neuropsychol Rev* 1995; 5(3): 203–221.
39. Jorgensen V, Opheim A, Halvarsson A, et al. Comparison of the Berg balance scale and the Mini-BESTest for assessing balance in ambulatory people with spinal cord injury: validation study. *Phys Ther* 2017; 97(6): 677–687.



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 ($\leq .43/1.0$), TUG (≥ 8.17 seconds), FSST (≥ 8.49 seconds), BBS ($\leq 50.5/56$), and ABC ($\leq 80.2/100$). 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.¹ 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.² 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

the probability with which an adverse event or condition may occur.

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,^{3,5,7,9,10} financial costs,¹¹ reduced mobility,^{12,13} and diminished quality of life.¹⁴ A major barrier to reducing falls among LLP users has been effective screening of those at risk.^{15,16} 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),¹⁷ timed up and go (TUG),¹⁸ four square step test (FSST),¹⁹ and activities-specific balance confidence (ABC) scale.²⁰ A number of studies have assessed the validity and reliability of these instruments among LLP users,^{15,16,21-27} but cutoff scores to infer fall risk have been established only for transtibial prosthesis users with limited prosthetic experience (ie, 6 months of prosthetic use).²¹ Clinicians, therefore, lack the necessary information to assess fall risk among individual patients within the larger LLP user population (ie, established transtibial and transfemoral prosthesis users).

Two different faller classifications, either ≥ 1 fall^{3,16,28,29} or ≥ 2 falls^{3,15,21} 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 ≥ 2 falls was used in a prior study.⁸ Adopting a ≥ 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, ≥ 1 fall classification). As a result, it may be easier to identify fallers versus non-fallers 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.^{16,21} 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,^{3,15,16,21} a secondary objective of this study was to determine whether cutoff scores and associated validity indices differed across two common fall classifications (ie, ≥ 1 fall vs ≥ 2 falls over the past 12 months). It was hypothesized that classifying fallers as ≥ 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 guideline³⁰ 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.³¹ 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).³² 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.^{1,33-35}

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).³⁶ 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?"³⁷⁻⁴⁰ 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^{3,16,28,29} and ≥ 2 falls^{3,15,21} over the past 12 months.

Clinical Balance Tests

Five clinical balance instruments, the ABC Scale,²⁰ the TUG,¹⁸ the FSST,¹⁹ the BBS,¹⁷ and the NBWT¹⁶ were administered and scored according to the developers' instructions (Appendix 1). Each balance instrument has demonstrated acceptable levels of validity^{15,16,21-23} and/or reliability (ie, intraclass correlation coefficient range .70-.99)^{15,22,23,25-27,41} 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^{33,42} 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.⁴³ The larger the AUC, the greater the test's general discriminative ability.⁴³ The AUC was selected as a summary measure of diagnostic accuracy.⁴⁴ The AUC can assume any value between 0 and 1, with a value of 0.5 representing chance,² and values greater than or equal to 0.8 recommended as the limit of clinical acceptability.⁴⁵ 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.⁴⁶ This point was identified by choosing the minimal value of the function $(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2$, a method that maximizes sensitivity and specificity.^{46,47} 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.² Values for sensitivity (ie, how often a diagnostic test is positive when a condition is present), and specificity (ie, how often a diagnostic test is negative in the absence of the condition) at that point were recorded.

One issue with sensitivity and specificity indices is that they have limited clinical utility.⁴⁸ Sensitivity, for example, does not describe how often patients with positive tests have the condition of interest (ie, a fall). Sensitivity indices, therefore, do not indicate the probability of the event or condition occurring,⁴⁹ information that is likely to be of greatest value to clinicians.¹ 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.⁵⁰ Likelihood ratios overcome these aforementioned limitations by quantifying how much the obtained test score increases or decreases the probability of an event occurring,^{1,35} 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.^{34,35,51,52} 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.⁴⁹ 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.¹ 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.⁵³ To avoid multicollinearity, any of the independent variables retained from the univariate logistic regression that had a strong correlation (ie, Spearman $\rho \geq .75$) and a lower odds ratio than the other independent variable with which it was correlated were excluded from the multivariate logistic regression model.⁵³ A multivariate logistic regression was then performed with all remaining independent variables included. Validity indices described previously were computed from the regression output. Cutoff 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.⁵² 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 fall 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

Fall Group	Age (y) Mean (SD)	Sex	Years since Amputation Mean (SD)	Amputation Level	Amputation Etiology	PLUS-M Mean (SD)	MFCL
No falls	44.9 (14.4)	M (11) F (5)	16.4 (12.8)	TT (14) TF (2)	Dysvascular (1) Nondysvascular (15)	58.5 (9.05)	K1 (0) K2 (4) K3 (10) K4 (2)
1 fall	41.4 (16.3)	M (2) F (5)	10.8 (4.8)	TT (3) TF (4)	Dysvascular (0) Nondysvascular (7)	56.2 (6.35)	K1 (0) K2 (1) K3 (2) K4 (4)
≥ 2 falls	55.4 (11.7)	M (8) F (9)	13.8 (14.8)	TT (8) TF (9)	Dysvascular (6) Nondysvascular (11)	49.2 (6.71)	K1 (2) K2 (10) K3 (5) K4 (0)
Total	48.7 (14.6)	M (21) F (19)	14.3 (12.6)	TT (25) TF (15)	Dysvascular (7) Nondysvascular (33)	54.1 (8.68)	K1 (2) K2 (15) K3 (17) K4 (6)

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.

Table 2

Clinical balance test scores stratified by fall history

Fall Status	NBWT (/1.0) Mean (SD) Range	TUG (s) Median (IQR) Range	FSST (s) Median (IQR) Range	BBS (/56) Median (IQR) Range	ABC (/100) Median (IQR) Range
No falls (n = 16)	0.51 (0.13) .27-.76	7.86 (2.83) 5.99-13.0	7.56 (4.87) 5.26-12.8	52.0 (8.50) 40.0-55.0	90.3 (18.9) 63.1-100
1 fall (n = 7)	0.53 (0.16) .25-.74	7.45 (3.04) 5.06-9.18	6.57 (3.08) 4.14-11.8	52.0 (6.00) 50.0-56.0	90.6 (14.1) 53.1-99.2
≥2 falls (n = 17)	0.25 (0.16)* .04-.63	10.0 (5.89) [†] 8.30-19.0	13.8 (6.76) [†] 6.27-39.3	45.0 (12.5) [†] 16.0-55.0	74 (26.9) 40.9-96.0
Total (n = 40)	0.41 (0.19) .04-.76	8.75 (3.97) 5.06-19.0	8.83 (5.94) 4.14-39.3	50.0 (8.50) 16.0-56.0	85.0 (23.7) 40.9-100

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.

*Test scores significantly different between ≥2 falls and 1 fall, as well as ≥2 falls and no falls $P < .004$ (One-way ANOVA).

[†]Test scores significantly different between ≥2 falls and 1 fall, as well as ≥2 falls and no falls $P < .027$ (Kruskal-Wallis Test of medians).

Table 3

Clinical balance test scores stratified by two faller classifications

≥1 fall (any falls)					
Fall Status	NBWT (/1.0) Mean (SD) Range	TUG (s) Median (IQR) Range	FSST (s) Median (IQR) Range	BBS (/56) Median (IQR) Range	ABC (/100) Median (IQR) Range
No falls (n = 16)	0.51 (.13) .27-.76	7.86 (2.83) 5.99-13.0	7.56 (4.87) 5.26-12.8	52.0 (8.50) 40.0-55.0	90.3 (18.9) 63.1-100
≥1 falls (n = 24)	0.33 (.20)* .04-.76	9.39 (5.87) 5.99-13.0	11.4 (8.54) 4.14-39.3	48.5 (10.8) 16.0-56.0	81.8 (24.9) 40.9-99.1
≥2 falls (multiple falls)					
Fall Status	NBWT (/1.0) Mean (SD) Range	TUG (s) Median (IQR) Range	FSST (s) Median (IQR) Range	BBS (/56) Median (IQR) Range	ABC (/100) Median (IQR) Range
0-1 falls (n = 23)	0.51 (.14) .25-.76	7.79 (1.66) 5.06-13.0	7.46 (3.75) 4.14-12.8	52.0 (7.00) 40.0-56.0	90.6 (15.6) 53.1-100
≥2 falls (n = 17)	0.25 (.16)* .04-.63	10.0 (5.89) [†] 8.30-19.0	13.8 (6.76) [†] 6.27-39.3	45.0 (12.5) [†] 16.0-56.0	74 (26.9) 40.9-96.0

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.

*Test scores significantly different between No falls and ≥1 fall, as well as ≥2 falls and 0-1 falls $P \leq .005$ (2-sided Independent t Test).

[†]Test scores significantly different between ≥2 falls and 0-1 fall $P \leq .001$ (Mann Whitney Test of medians).

Table 4

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

≥1 fall (any falls)						
Test	AUC (95% CI)	Cutoff Score	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
NBWT (/1.00)	0.81 (.62-.91)	≤0.43	73% (53%-90%)	76% (54%-96%)	3.0 (1.5-6.9)	0.36 (.19-.77)
TUG (s)	0.71 (.54-.88)	≥8.17	83% (68%-98%)	68% (46%-92%)	2.6 (1.3-5.6)	0.24 (.13-.56)
FSST (s)	0.70 (.53-.86)	≥8.49	74% (58%-92%)	68% (46%-92%)	2.4 (1.1-5.2)	0.36 (.17-.78)
BBS (/56)	0.66 (.47-.83)	≤50.5	67% (48%-86%)	62% (39%-86%)	1.8 (.89-3.6)	0.53 (.27-1.1)
ABC (/100)	0.65 (.47-.82)	≤80.2	50% (30%-70%)	74% (54%-96%)	1.9 (.78-5.1)	0.67 (.41-1.1)
Model 1	0.84 (.71-.96)	N/A	80% (64%-96%)	73% (51%-96%)	3.0 (1.3-7.1)	0.27 (.12-.63)
≥2 falls (multiple falls)						
Test	AUC (95% CI)	Cutoff Score	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
NBWT (/1.00)	0.89 (.78-1.0)	≤0.43	88% (73%-100%)	79% (65%-92%)	4.2 (1.7-6.9)	0.16 (.042-.60)
TUG (s)	0.88 (.78-.98)	≥9.25	82% (64%-100%)	83% (67%-98%)	4.7 (1.9-11.9)	0.21 (.075-.61)
FSST (s)	0.88 (.76-.99)	≥8.71	94% (83%-100%)	74% (60%-92%)	3.6 (1.8-7.3)	0.08 (.012-.54)
BBS (/56)	0.81 (.66-.92)	≤50.5	88% (73%-100%)	70% (51%-88%)	2.9 (1.5-5.5)	0.17 (.045-.64)
ABC (/100)	0.71 (.54-.87)	≤80.2	65% (42%-87%)	78% (61%-95%)	3.0 (1.3-7.0)	0.45 (.23-.89)
Model 2	0.90 (.79-1.0)	N/A	76% (56%-97%)	83% (67%-98%)	4.4 (1.7-11.3)	0.28 (.12-.68)

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.

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),^{24,54-58} 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 TUG^{16,22} 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 complementary 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.^{3,5,7,9,11} 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 falls-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.⁸ 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.⁶⁰⁻⁶⁴ 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,^{15,21,65} 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).¹ 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%),^{24,54-58} potentially limiting the generalization of study results. Although this is a challenge in most,^{15,23,25,66,67} but not all^{8,21,68-70} 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),^{54-56,58} 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 frequency⁷¹ and lend itself to recall bias.^{72,73} 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.⁷⁴

Fall-related injuries were not recorded. Determining whether multiple versus nonmultiple fallers are more likely to suffer a fall-related injury,^{3,10} 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,⁷⁵ protective stepping,⁷⁶ number of medications,⁹ and sense of vibration⁷⁷ 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

- Riddle DL, Stratford PW. Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. *Phys Ther*. 1999;79(10):939-948.
- Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*. 2006;163(7):670-675.
- Wong C, Chihuri S, Li G. Risk of fall-related injury in people with lower limb amputations: a prospective cohort study. *J Rehabil Med*. 2016;48(1):80-85.
- Kim J, Major MJ, Hafner B, Sawers A. Frequency and circumstances of falls reported by ambulatory unilateral lower limb prosthesis users: a secondary analysis. *PM R*. 2019;11(4):344-353.
- Kulkarni J, Wright S, Toole C, Morris J, Hirons R. Falls in patients with lower limb amputations: prevalence and contributing factors. *Physiotherapy*. 1996;82(2):130-136.
- Gauthier-Gagnon C, Grise MC, Potvin D. Enabling factors related to prosthetic use by people with transtibial and transfemoral amputation. *Arch Phys Med Rehabil*. 1999;80(6):706-713.
- Miller WC, Speechley M, Deathe B. The prevalence and risk factors of falling and fear of falling among lower extremity amputees. *Arch Phys Med Rehabil*. 2001;82(8):1031-1037.
- Wong CK, Chen CC, Blackwell WM, Rahal RT, Benoy SA. Balance ability measured with the Berg balance scale: a determinant of fall history in community-dwelling adults with leg amputation. *J Rehabil Med*. 2015;47(1):80-86.
- Pauley T, Devlin M, Heslin K. Falls sustained during inpatient rehabilitation after lower limb amputation: prevalence and predictors. *Am J Phys Med Rehabil*. 2006;85(6):521-532.
- Chihuri S, Wong CK. Factors associated with the likelihood of fall-related injury among people with lower limb loss. *Inj Epidemiol*. 2018;5(1):42-50.
- Mundell B, Maradit Kremers H, Visscher S, Hoppe K, Kaufman K. Direct medical costs of accidental falls for adults with transfemoral amputations. *Prosthet Orthot Int*. 2017;41(6):564-570.
- van Velzen JM, van Bennekom CA, Polomski W, Slootman JR, van der Woude LH, Houdijk H. Physical capacity and walking ability after lower limb amputation: a systematic review. *Clin Rehabil*. 2006;20(11):999-1016.
- Miller WC, Deathe AB, Speechley M, Koval J. The influence of falling, fear of falling, and balance confidence on prosthetic mobility and social activity among individuals with a lower extremity amputation. *Arch Phys Med Rehabil*. 2001;82(9):1238-1244.
- Pezzin LE, Dillingham TR, MacKenzie EJ. Rehabilitation and the long-term outcomes of persons with trauma-related amputations. *Arch Phys Med Rehabil*. 2000;81(3):292-300.
- Major MJ, Fatone S, Roth EJ. Validity and reliability of the Berg Balance Scale for community-dwelling persons with lower-limb amputation. *Arch Phys Med Rehabil*. 2013;94(11):2194-2202.
- Sawers A, Hafner B. Validation of the narrowing beam walking test in lower limb prosthesis users. *Arch Phys Med Rehabil*. 2018;99(8):1491-1498.
- Berg K, Wooddauphinee S, Williams JI, Maki B, Gayton D. Measuring balance in the elderly: validation of an instrument. *J Am Geriatr Soc*. 1991;39(8):A65-A65.
- Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc*. 1991;39(2):142-148.
- Dite W, Temple VA. A clinical test of stepping and change of direction to identify multiple falling older adults. *Arch Phys Med Rehabil*. 2002;83(11):1566-1571.
- Powell LE, Myers AM. The activities-specific balance confidence (ABC) scale. *J Gerontol A Biol Sci Med Sci*. 1995;50A(1):M28-M34.
- Dite W, Connor HJ, Curtis HC. Clinical identification of multiple fall risk early after unilateral transtibial amputation. *Arch Phys Med Rehabil*. 2007;88(1):109-114.
- Schoppen T, Boonstra A, Groothoff JW. The timed "up and go" test: reliability and validity in persons with unilateral lower limb amputation. *Arch Phys Med Rehabil*. 1999;80(7):825-828.
- Miller WC, Deathe AB, Speechley M. Psychometric properties of the activities-specific balance confidence scale among individuals with a lower-limb amputation. *Arch Phys Med Rehabil*. 2003;84(5):656-661.
- Hafner BJ, Morgan SJ, Askew RL, Salem R. Psychometric evaluation of self-report outcome measures for prosthetic applications. *J Rehabil Res Dev*. 2016;53(6):797-812.
- Cardoso JR, Beisheim EH, Horne JR, Sions JM. Test-retest reliability of dynamic balance performance-based measures among adults with a unilateral lower-limb amputation. *PM R*. 2019;11(3):243-251.
- Resnik L, Borgia M. Reliability of outcome measures for people with lower-limb amputations: distinguishing true change from statistical error. *Phys Ther*. 2011;91(4):555-565.
- Wong CK. Interrater reliability of the Berg balance scale when used by clinicians of various experience levels to assess people with lower limb amputations. *Phys Ther*. 2014;94(3):371-378.
- Berg WP, Alessio HM, Mills EM, Tong C. Circumstances and consequences of falls in independent community-dwelling older adults. *Age Ageing*. 1997;26(4):261-268.
- Topper AK, Maki BE, Holliday PJ. Are activity-based assessments of balance and gait in the elderly predictive of risk of falling and/or type of fall? *J Am Geriatr Soc*. 1993;41(5):479-487.
- Elm v E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296.
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.
- Hanspal RS, Fisher K, Nieveen R. Prosthetic socket fit comfort score. *Disabil Rehabil*. 2003;25(22):1278-1280.
- Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 1994;309(6948):188.
- Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329(7458):168-169.
- Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. *BMJ*. 1997;315(7107):540-543.
- Hafner BJ, Gaunard IA, Morgan SJ, Amtmann D, Salem R, Gailey RS. Construct validity of the prosthetic limb users survey of mobility (PLUS-M) in adults with lower limb amputation. *Arch Phys Med Rehabil*. 2016;98(2):277-285.
- Askham J, Glucksman E, Owens P, Swift C. *Home and Leisure Accident Research: A Review of Research on Falls among Elderly People*. London, UK: Age Concern Institute of Gerontology, King's College London, and Department of Trade and Industry; 1990.
- Lord SR, Ward JA, Williams P, Anstey KJ. Physiological factors associated with falls in older community-dwelling women. *J Am Geriatr Soc*. 1994;42(10):1110-1117.
- Lamb SE, Jørstad-Stein EC, Hauer K, Becker C. Development of a common outcome data set for fall injury prevention trials: the prevention of falls network Europe consensus. *J Am Geriatr Soc*. 2005; 53:1618-1622.
- Hauer K, Lamb SE, Jørstad EC, Todd C, Becker C. Systematic review of definitions and methods of measuring falls in randomized controlled fall prevention trials. *Age Ageing*. 2006;35(1):5-10.
- Clemens SM, Gailey RS, Bennett CL, Pasquina PF, Kirk-Sanchez NJ, Gaunard IA. The Component Timed-Up-and-Go test: the utility and psychometric properties of using a mobile application to determine prosthetic mobility in people with lower limb amputations. *Clin Rehabil*. 2018;32(3):388-397.
- Portney LG, Watkins MP. *Foundations of Clinical Research*. Upper Saddle River, NJ: Pearson; 2009.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic curve. *Radiology*. 1982;143(1):29-36.

44. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.
45. Haynes RB. *Clinical Epidemiology: how we Do Clinical Practice Research*. Philadelphia, PA: Lippincott, Williams, and Wilkins; 2006.
46. Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr*. 2007;96(5):644-647.
47. Schlenstedt C, Brombacher S, Hartwigsen G, Weisser B, Moeller B, Deuschl G. Comparison of the Fullerton advanced balance scale, mini-BESTest, and Berg balance scale to predict falls in Parkinson's disease. *Phys Ther*. 2016;96(4):494-501.
48. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Philadelphia, PA: Lippincott, Williams, and Wilkins; 1991.
49. Altman DG, Bland JM. Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ*. 1994;308(6943):1552-1552.
50. Altman DG, Bland JM. Statistics notes: diagnostic tests 2: predictive values. *BMJ*. 1994;309(6947):102-102.
51. Crombie DL. Diagnostic process. *J Coll Gen Pract*. 1963;6:579-589.
52. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
53. Paul SS, Canning CG, Sherrington C, Lord SR, Close JCT, Fung VSC. Three simple clinical tests to accurately predict falls in people with Parkinson's disease. *Mov Disord*. 2013;28(5):655-662.
54. Ziegler-Graham K, MacKenzie EJ, Ephraim PL, Trivison TG, Brookmeyer R. Estimating the prevalence of limb loss in the United States: 2005 to 2050. *Arch Phys Med Rehabil*. 2008;89(3):422-429.
55. Pezzin LE, Dillingham TR, MacKenzie EJ, Ephraim P, Rossbach P. Use and satisfaction with prosthetic limb devices and related services. *Arch Phys Med Rehabil*. 2004;85(5):723-729.
56. Ehde DM, Czerniecki JM, Smith DG, et al. Chronic phantom sensations, phantom pain, residual limb pain, and other regional pain after lower limb amputation. *Arch Phys Med Rehabil*. 2000;81(8):1039-1044.
57. American Board for Certification in Orthotics, Prosthetics, & Pedorthics (ABC). *Practice Analysis of Certified Practitioners in the Disciplines of Orthotics and Prosthetics*. Alexandria, VA; 2015:1-60.
58. Wurdeman SR, Stevens PM, Campbell JH. Mobility analysis of Amputees II: comorbidities and mobility in lower limb prosthesis users. *Am J Phys Med Rehabil*. 2018;97(11):782-788.
59. Hunter SW, Batchelor F, Hill KD, Hill A-M, Mackintosh S, Payne M. Risk factors for falls in people with a lower limb amputation: a systematic review. *PM R*. 2017;9(2):170-180.
60. Bowman J. Challenges to measuring outcomes in occupational therapy: a qualitative focus group study. *Br J Occup Ther*. 2006;69(10):464-472.
61. Gaunard I, Spaulding SE, Amtmann D, et al. Use of and confidence in administering outcome measures among clinical prosthetists: results from a national survey and mixed-methods training program. *Prosthet Orthot Int*. 2015;39(4):314-321.
62. Jette DU, Halbert J, Iverson C, Miceli E. Use of standardized outcome measures in physical therapist practice: perceptions and applications. *Phys Ther*. 2009;89:125-135.
63. Duncan EAS, Murray J. The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Serv Res*. 2012;12(1):96-104.
64. Valier A, Jennings AL, Parsons JT. Benefits of and barriers to using patient-rated outcome measures in athletic training. *J Athl Train*. 2014;49(5):674-683.
65. Jorgensen V, Opheim A, Halvarsson A, Franzén E, Roaldsen KS. Comparison of the Berg balance scale and the mini-BESTest for assessing balance in ambulatory people with spinal cord injury: validation study. *Phys Ther*. 2017;97(6):677-687.
66. Reid L, Thomson P, Besemann M, Dudek N. Going places: does the two-minute walk test predict the six-minute walk test in lower extremity amputees? *J Rehabil Med*. 2015;47(3):256-261.
67. Lin S-J, Bose NH. Six-minute walk test in persons with transtibial amputation. *Arch Phys Med Rehabil*. 2008;89(12):2354-2359.
68. Sakakibara BM, Miller WC, Backman CL. Rasch analyses of the activities-specific balance confidence scale with individuals 50 years and older with lower-limb amputations. *Arch Phys Med Rehabil*. 2011;92(8):1257-1263.
69. Brooks D, Hunter JP, Parsons J, Livsey E, Quirt J, Devlin M. Reliability of the two-minute walk test in individuals with transtibial amputation. *Arch Phys Med Rehabil*. 2002;83(11):1562-1565.
70. Mandel A, Paul K, et al. Balance confidence and activity of community-dwelling patients with transtibial amputation. *J Rehabil Res Dev*. 2016;53(5):551-560.
71. Hill K, Schwarz J, Flicker L, Carroll S. Falls among healthy, community-dwelling, older women: a prospective study of frequency, circumstances, consequences and prediction accuracy. *Aust N Z J Public Health*. 1999;23(1):41-48.
72. Cummings SR, Nevitt MC, Kidd S. Forgetting falls. The limited accuracy of recall of falls in the elderly. *J Am Geriatr Soc*. 1988;36(7):613-616.
73. Feldman F, Robinovitch SN. Recalling the mechanics of falls: Young adults cannot accurately describe the sites of impact immediately after a fall occurs. *Osteoporos Int*. 2006;17(6):955-955.
74. Muir SW, Speechley M. Establishing predictive validity of the Fullerton advanced balance scale. *Arch Phys Med Rehabil*. 2010;91(7):1147-1148.
75. Vanicek N, Strike S, McNaughton L, Polman R. Gait patterns in transtibial amputee fallers vs. non-fallers: biomechanical differences during level walking. *Gait Posture*. 2009;29(3):415-420.
76. Crenshaw JR, Kaufman KR, Grabiner MD. Compensatory-step training of healthy, mobile people with unilateral, transfemoral or knee disarticulation amputations: a potential intervention for trip-related falls. *Gait Posture*. 2013;38(3):500-506.
77. Quai TM, Brauer SG, Nitz JC. Somatosensation, circulation and stance balance in elderly dysvascular transtibial amputees. *Clin Rehabil*. 2005;19(6):668-676.

Disclosure

A.S. Department of Kinesiology and Nutrition, University of Illinois at Chicago, Chicago, IL. Address correspondence to: A. S. Department of Kinesiology and Nutrition, University of Illinois at Chicago, 1919 W Taylor Street, Chicago, IL 60612, USA. Tel: 312-996-1427, Fax: 312-413-0319, e-mail: asawers@uic.edu
Disclosure: none

B.J.H. Department of Rehabilitation Medicine, University of Washington, Seattle, WA
Disclosure: none

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 National Institutes of Health (NIH) under award number K12HD073945 (AS), and the Orthotics and Prosthetics Education and Research Foundation, Inc. (OPERF) under grant number OPERF-SGA-2016-1 (AS, BH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the DoD, NIH, or OPERF.

Submitted for publication January 10, 2019; accepted March 18, 2019.

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

Journal:	<i>Physical Therapy</i>
Manuscript ID	PTJ-2019-0247.R1
Manuscript Category:	Original Research - Measurement Property Evaluation
Section:	Balance and Falls
Keywords:	Amputation, Outcome Assessment (Health Care), Reliability, Accidental Falls, Reproducibility of Results

SCHOLARONE™
Manuscripts

1
2 1 **Running Head:** Reliability of balance tests in amputees

3
4 2
5
6 3 **Title:** Inter-rater and test-retest reliability of performance-based clinical balance tests administered to
7
8
9 4 established unilateral lower limb prosthesis users

10
11 5
12
13 6 **Prior Publication:**
14
15 7 The work described in this manuscript has not been previously published, nor is it under consideration
16
17
18 8 for publication elsewhere.

19
20 9
21
22 10 **Financial Disclosure:** The authors declare no competing interest.

23
24
25 11
26
27 12 **Institutional Review:** All study procedures were reviewed and approved by a XXXX institutional
28
29 13 review board.

30
31
32 14
33
34 15 **Clinical Trials Registration:** Not applicable

35
36 16
37
38
39 17
40
41 18
42
43 19
44
45 20
46
47
48 21
49
50 22
51
52 23
53
54
55 24
56
57 25

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., $\geq .98$). 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., $ICC \geq .70$ and $ICC \geq .90$, 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.

1
2 51 **Key Words:** amputation, accidental falls, outcome assessment, reliability, reproducibility of results.
3
4 52 3696/4000 words (introduction through conclusion)
5
6 53
7
8
9 54 **Abbreviations:**
10
11 55 10mWT = 10 meter Walk Test
12
13 56 ANOVA = Analysis of Variance
14
15 57 CCI = Charlson Comorbidity Index
16
17
18 58 COSMIN = COnsensus-based Standards for the selection of health Measurement INstruments
19
20 59 FSST = Four Square Step Test
21
22 60 ICC = intraclass correlation coefficient
23
24
25 61 LLP = lower limb prosthesis
26
27 62 MDC = minimal detectable change
28
29 63 MFCL = Medicare Functional Classification Level
30
31
32 64 NBWT = Narrowing Beam Walking Test
33
34 65 PLUS-M = Prosthetic Limb Users Survey of Mobility
35
36 66 REDCap = Research Electronic Data Capture
37
38
39 67 SCS = Socket Comfort Score
40
41 68 SD = standard deviation
42
43 69 SEM = standard error of measurement
44
45 70 STROBE: Strengthening The Reporting of Observational studies in Epidemiology
46
47
48 71 TUG = Timed Up and Go
49
50 72
51
52 73
53
54
55 74
56
57 75
58
59
60

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) users¹⁻³. Despite the demand for robust documentation, several barriers inhibit regular use of standardized outcome measures in clinical practice²⁻⁶. 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 falls¹⁰⁻¹⁴, and their related consequences¹⁵⁻¹⁷ 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 greater²⁵⁻²⁸. 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.

1
2 101
3
4 102 The SEM reflects the error associated with a *single* measurement¹⁹. The SEM can therefore be used
5
6 103 to estimate the precision of an individual's score on a test^{19,29,30}, as it can be used to derive confidence
7
8 104 intervals (CIs) around a score. The MDC is an estimate of the amount of change needed to exceed the
9
10
11 105 error associated with two measurements on the same test within the same individual³¹. An individual
12
13 106 patient is considered to have changed only when the difference between the previous score and the
14
15 107 current score exceeds the MDC of the test³². The SEM and MDC can therefore help clinicians interpret
16
17
18 108 test scores among individual patients by monitoring whether there is a real difference in their health or
19
20 109 performance with respect to a reference value (i.e., using SEM), and/or whether their health or
21
22 110 performance has changed between testing sessions (i.e., using MDC)¹⁹.

23
24
25 111
26
27 112 Few performance-based clinical ~~balance~~ tests with evidence of validity among LLP users have been
28
29 113 assessed for evidence of reproducibility³³⁻³⁶. **In the absence of reliability indices like test-retest ICCs**
30
31
32 114 **or SEM and MDC values, clinicians cannot select tests suitable for individual level decision-**
33
34 115 **making, or interpret test scores of individual LLP users in clinical practice.** Efforts to study the
35
36 116 reproducibility of performance-based clinical ~~balance~~ tests administered to LLP users have been limited
37
38
39 117 by several factors. First, sample sizes are generally small ³³⁻³⁵, and often below recommendations for
40
41 118 establishing reliability indices³⁷. Second, evidence of reproducibility for some tests is dated and may not
42
43 119 reflect best practices for assessing test-retest reliability³⁵. Lastly, study methodology (e.g., selection of
44
45 120 raters and corresponding analysis) may limit the generalizability of study results to the broader
46
47
48 121 population of LLP users and test administrators^{34,36}. Additionally, a number of common performance-
49
50 122 based clinical ~~balance~~ tests (e.g., FSST, 10mWT, TUG) have not yet been assessed for inter-rater
51
52 123 reliability in LLP users. **Many performance-based tests also evaluate unique aspects of balance and**
53
54
55 124 **walking (i.e., straight ahead walking (10MWT), turns (TUG), balancing along beam (NBWT), and**
56
57 125 **multi-planar stepping (FSST)). As such, it is important to evaluate each test and determine if it**
58
59
60

and its components are suitable for clinical practice. Finally, given the relative nature of that test-retest reliability **is relative to the sample of participants from which it is derived³⁸, 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., $ICC \geq 0.7$) and individual decision-making (i.e., $ICC \geq 0.9$) 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)³⁹, Timed Up and Go (TUG)⁴⁰, Four Square Step Test (FSST)⁴¹, and the 10 meter Walk Test (10mWT)⁴² when administered to LLP users. Best practices in conducting observational studies (i.e., STROBE: Strengthening The Reporting of Observational studies in Epidemiology⁴³), as well as the collection and

1
2 151 reporting of reliability data (i.e., COnsensus-based Standards for the selection of health Measurement
3
4 152 INstruments (COSMIN))^{19,20,37} were followed during the study. All data were stored and managed using
5
6 153 a REDCap database hosted at XXXX⁴⁴. Study protocols were reviewed and approved by institutional
7
8
9 154 review boards at each site. All individuals provided written consent prior to participation.
10

11 155
12
13 156 **Participants**

14
15 157 **Study flyers were used to recruit** LLP users from local prosthetic clinics in XXXX and XXXX. A
16
17
18 158 convenience sampling method was used. Inclusion criteria were 18 years of age or older; unilateral
19
20 159 amputation at the transtibial or transfemoral level; amputation due to trauma, dysvascular complications,
21
22 160 tumor, or infection; use of a prosthesis for at least one year (**i.e., established lower limb prosthesis**
23
24
25 161 **users**); use of a comfortable prosthesis to ambulate; **able to walk 10 meters without an upper**
26
27 162 **extremity mobility aid (e.g., cane)**; and able to read, write, and speak English. Participants were
28
29 163 excluded from participation if they had an amputation of a second limb (e.g., arm or contralateral leg);
30
31
32 164 contralateral complications (e.g., hip replacement, ulcers or infections), advanced neurologic disorder
33
34 165 (**e.g., Stroke, Multiple Sclerosis, Parkinson’s disease**); or severe cardiopulmonary disease that would
35
36 166 prevent them from completing study procedures.
37

38
39 167
40
41 168 An a priori sample size of 38 LLP users was estimated based on a desired 95% confidence interval
42
43 169 with a width greater than or equal to .10 around an ICC of .80⁴⁵. Psychometric standards, however,
44
45 170 recommend a minimum sample of 50 participants to estimate reliability statistics¹⁹. Based on the more
46
47
48 171 conservative estimate (i.e., recommended psychometric standards), and planning for modest attrition
49
50 172 (i.e., 20%) between test and retest sessions^{46,47}, a sample of 60 LLP users was targeted.
51

52 173
53
54
55 174 **Procedures**
56
57
58
59
60

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 MDC₉₀ of the SCS in LLP users (i.e., ± 2.82 points)⁵².

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

The TUG⁴⁰, FSST⁴¹ 10mWT⁴², and the NBWT³⁹ 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

1
2 200 been designed or used to assess fall risk in other populations, and has evidence of validity among LLP
3
4 201 users^{35,39,53,54}.

5
6 202
7
8
9 203 Standardized protocols were established to ensure fidelity in test administration and scoring between
10
11 204 study sites and raters. All raters were instructed in test procedures using a common set of training
12
13 205 documents. All raters practiced administration of the tests under the supervision of study (or site)
14
15 206 principal investigators. Test equipment was purchased or fabricated using explicit written instructions,
16
17
18 207 and verified with photographs and videos during a training session. Verbal instructions were
19
20 208 standardized to help ensure consistency in data collection across **the two** study sites⁵⁵. Further, no
21
22 209 **communication was allowed between concurrent raters during testing to limit** their awareness of
23
24
25 210 each other's assessments, which could influence scoring⁵⁶. Finally, data collected on standardized paper
26
27 211 forms were double entered by research staff to minimize data-entry errors⁵⁷. **Participants were**
28
29 212 **provided two minutes rest between tests, more if requested.**

30
31
32 213
33
34 214 **A small pool of qualified raters was consistently used at each study site. Specifically, across**
35
36 215 **study sites** raters included two researchers (i.e., PhD), three PhD students (one certified prosthetist-
37
38
39 216 orthotist, one physical therapist, and one bioengineer), and three prosthetic and orthotic graduate
40
41 217 students. **These raters were selected because of their qualifications, training, and knowledge of the**
42
43 218 **instruments under investigation. Further, these qualities reflected those of** individuals working in a
44
45
46 219 clinical or research setting, thereby contributing to the generalizability of the subsequent reliability
47
48 220 indices²⁰. **Testing during session two was performed by one of the raters from session one.**

49
50 221
51
52 222 **Data Analysis**

53
54
55 223 Continuous variables were evaluated for departures from normality using the Shapiro-Wilk test⁵⁸.
56
57 224 Measures of central tendency and dispersion were calculated for each variable based on their
58
59
60

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-distribution⁶⁰. 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 comparisons²⁵⁻²⁸, and .90 or greater for individual-level decisions^{19,25-28}).

Standard error of measurement (SEM) estimates were calculated as $SEM = SD * \sqrt{(1 - ICC)}$, 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 session³³, 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 $MDC_{90} = 1.64 * \sqrt{SEM} * \sqrt{2}$ where 1.64 is the z-score for a 90% confidence interval³⁸.

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,

1
2 250 diabetes, HIV/AIDS), respectively. Observed test scores (Supplement 1) for the FSST, TUG, and
3
4 251 10mWT were non-normally distributed (Shapiro-Wilk $W=.798-.919$, $p\leq.001$). In contrast, NBWT scores
5
6 252 were normally distributed (Shapiro-Wilk $W=.957-.974$, $p\geq.195$). Despite departures from normality,
7
8
9 253 parametric 2-sided t-tests were used to compare test scores between rater and across test sessions
10
11 254 because of their robustness to withstand violations of normality in larger samples (i.e., $n>30$)⁶¹.
12

13 255
14
15 256 *Inter-rater Reliability*
16

17
18 257 Two-sided paired t-tests revealed no statistically significant differences in mean test scores between
19
20 258 raters for the TUG ($t=1.609$, $p=.113$), FSST ($t=-.753$, $p=.455$), NBWT ($t=1.276$, $p=.207$), or 10mWT
21
22 259 ($t=-.518$, $p=.607$) (Table 2). Inter-rater reliability ICCs (model 1,1) were high (i.e., $\geq.98$) for all clinical
23
24
25 260 tests (Table 2).
26

27 261
28
29 262 *Test-retest Reliability*
30

31
32 263 Retest sessions occurred at a mean of 6.7 days (SD = 2.3 day) after the initial testing session. No
33
34 264 changes in health, prosthetic prescription, prosthesis function, or SCS (test session 1: 7.62 (SD=1.97),
35
36 265 test session 2: 7.43 (SD=1.61), $t=1.026$, $p=.309$) were reported between test sessions by study
37
38
39 266 participants, as a group or individually. Statistically significant differences in test scores were observed
40
41 267 between test sessions for the TUG ($t=2.968$, $p=.004$), but not the 10mWT ($t=2.315$, $p=.024$), FSST
42
43 268 ($t=.821$, $p=.415$), or NBWT ($t=-1.821$, $p=.066$) (Table 3). Test-retest ICCs (model 2,1) varied by test,
44
45
46 269 ranging from .88 for the TUG, to .97 for the FSST (Table 3).
47

48 270
49
50 271 *Standard Error of Measurement and Minimal Detectable Change*
51

52 272 SEM and MDC₉₀ varied between .39 - .96, and .91s - 2.2s, respectively, among the time-based tests
53
54
55 273 (i.e., 10mWT, TUG, and FSST) (Table 3). Estimates of SEM and MDC₉₀ for the NBWT, a test of
56
57 274 normalized walking distance, were .07 and .16, respectively.
58
59
60

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., $ICC \geq .70$), while the FSST, 10mWT, and NBWT are also suitable for individual-level applications among established unilateral LLP users (i.e., $ICC \geq .90$). **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., $ICC \geq 0.7$) and individual decision-making (i.e., $ICC \geq 0.9$).** 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., MDC_{90}). 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

1
2 300 previously reported for the 10mWT (i.e., .93-.98)⁶⁶, as well as the TUG, as described by Resnik et al.,
3
4 301 (2011) (i.e., .88 [95% CI: .80-.94]), but not Clemens et al., (2018) (i.e., .98 [95% CI: .97-.99]) or
5
6 302 Jayakaran et al., (2011) (i.e., .95 [95% CI: .85-.98]). The larger TUG test-retest ICC reported by
7
8 303 Clemens et al., (2018) may be explained by their experimental design. Calculating the test-retest ICC
9
10
11 304 using back-to-back trials of the TUG, rather than trials administered several days apart^{33,38}, excluded
12
13 305 session-to-session variation within subjects, an important component of error when calculating test-
14
15 306 retest ICCs³⁸. Excluding session-to-session variation likely increases the magnitude of test-retest ICCs
16
17
18 307 and attenuates the corresponding indices of measurement error. The test-retest ICC for the TUG reported
19
20 308 by Jayakaran et al., (2011) may have been affected by their atypical scoring (i.e., the mean of two trials
21
22 309 rather than the conventional “best” of two trials)⁴⁰. Using the mean may have stabilized the score,
23
24
25 310 reduced error, and increased reliability. To date, evidence of test-retest reliability for the FSST has been
26
27 311 limited to other patient populations (e.g., Parkinson’s disease, multiple sclerosis, post-stroke)⁶⁷, but has
28
29 312 yet to be reported among LLP users.

31
32 313
33
34 314 Notable departures from previously reported values of the standard error of measurement (SEM)
35
36 315 were found for the studied tests. The SEM for the TUG in the present study (i.e., .96) was larger than
37
38 316 one previously reported value, .55³⁶, and smaller than another, 1.6³³. Since the SEM is an arithmetic
39
40
41 317 function of the test-retest ICC and the standard deviation of test scores from a study sample, the
42
43 318 differences can be attributed to the notably larger test-retest ICC reported by Clemens et al., (2018), and
44
45 319 the nearly two-fold increase in between-subjects variability reported by Resnik et al., (2011),
46
47
48 320 respectively⁶⁸. To date, the SEM for the 10mWT or FSST among LLP users has not been reported.

49
50 321
51
52 322 Point estimates for the MDC₉₀ also differed from previously reported values^{33,36}. Given that MDC is
53
54
55 323 a scalar product of the SEM, these differences can be explained by the earlier discussion of observed
56
57 324 differences in SEM. Despite the disparities in MDC₉₀ point estimates in this study and those reported

previously, the 95% CIs of MDC₉₀ 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 tests³⁸. 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., $ICC \geq .70$)²⁵⁻²⁸, and/or individual-level applications (i.e., $ICC \geq .90$)^{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 coefficient⁶⁹, determines if reliability results can be generalized beyond the participants and raters included in the study³⁸. Consistent with most reproducibility studies conducted to-date^{33,34,36}, test-retest ICCs in the present study were calculated using ICC (2,1), which includes both random and systematic error^{22,38}. Model 2 permits the generalization of reliability results to other raters with similar participants or patients³⁸. 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^{29,38}. 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)^{70,71} are above and beyond changes that may be attributable to measurement error³¹. 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 MDC₉₀ 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

1
2 400 **cause of amputation between the two studies. Additional research is however required to verify**
3
4 401 **these between-session differences and offer potential explanations.** Regardless, these “time effects”
5
6 402 suggest that practice effects may persist between test sessions for the TUG when it is applied with LLP
7
8 403 users^{38,72,73}. **While not evaluated in the present study, within-session trial-to-trial practice effects**
9
10
11 404 **may also exist among a host of performance-based clinical tests. These practice effects may limit**
12
13 405 **not only the reliability of performance-based tests, but also their validity (i.e., diagnostic accuracy)**
14
15 406 ⁷³. **Previous research on the NBWT has shown that changes to administration and scoring**
16
17
18 407 **procedures (i.e., providing additional practice trials) can accommodate practice effects and**
19
20 408 **improve psychometric performance**⁷³. ~~It is possible that changes to TUG administration and scoring~~
21
22 409 ~~procedures may be able to resolve these issues~~⁷³. ~~It may therefore be prudent to examine performance-~~
23
24
25 410 ~~based clinical balance tests’ administration and scoring procedures to determine if they are susceptible to~~
26
27 411 ~~practice effects or other sources of systematic error~~³⁸. **If they are, modifications to administration and**
28
29 412 **scoring procedures may prove effective in accommodating within and between session practice**
30
31
32 413 **effects.**

36 415 **Conclusion**

38
39 416 Results of this study provide valuable evidence of inter-rater and test-retest reliability, measurement
40
41 417 error, and minimal detectable change for performance-based clinical balance tests among established
42
43 418 unilateral LLP users. **These reliability metrics provide clinicians with actionable evidence they can**
44
45 419 **use to select tests for use with individual patients, interpret the resultant test scores, and**
46
47
48 420 **effectively use the test results to make informed clinical decisions.** All studied tests demonstrated
49
50 421 sufficient reliability to be considered appropriate for group-level comparisons (i.e., ICC ≥ .70), while the
51
52 422 FSST, 10mWT, and NBWT were also found to be suitable for individual-level applications among
53
54
55 423 established unilateral LLP users (i.e., ICC ≥ .90). SEM and MDC estimates of the studied tests provide
56
57 424 clinicians and researchers with the information necessary to assess differences or changes in balance
58
59
60

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

1. Heinemann AW, Connelly L, Ehrlich-Jones L, Fatone S. Outcome instruments for prosthetics: clinical applications. *Phys Med Rehabil Clin N Am*. 2014;25(1):179-198.
2. Wedge FM, Braswell-Christy J, Brown CJ, Foley KT, Graham C, Shaw S. Factors influencing the use of outcome measures in physical therapy practice. *Physiother Theory Pract*. 2012;28(2):119-133.
3. Jette DU, Halbert J, Iverson C, Miceli E. Use of standardized outcome measures in physical therapist practice: perceptions and applications. *Phys Ther*. 2009;89(2):125-135.
4. Bowman J. Challenges to measuring outcomes in occupational therapy: A qualitative focus group study. *Br J Occup Ther*. 2006;69(10):464-472.
5. Duncan EAS, Murray J. The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Serv Res*. 2012;12(1):96-104.
6. Gaunaud I, Spaulding SE, Amtmann D, et al. Use of and confidence in administering outcome measures among clinical prosthetists: Results from a national survey and mixed-methods training program. *Prosthet Orthot Int*. 2015;39(4):314-321.
7. Hafner BJ, Spaulding SE, Salem R, Morgan SJ, Gaunaud I, Gailey R. Prosthetists' perceptions and use of outcome measures in clinical practice: Long-term effects of focused continuing education. *Prosthet Orthot Int*. 2016;41(3):1-8.
8. Stapleton T, McBrearty C. Use of Standardized assessments and outcome measures among a sample of Irish occupational therapists working with adults with physical disabilities. *Br J Occup Ther*. 2009;72(2):55-64.
9. Belazi D, Goldfarb NI, He H. Measuring health-related quality of life in the clinical setting. *Expert Rev Pharmacoecon Outcomes Res*. 2014;2(2):109-117.
10. Kim J, Major MJ, Hafner B, Sawers A. Frequency and circumstances of falls reported by ambulatory unilateral lower limb prosthesis users: A secondary analysis. *PM R*. 2019;11(4):344-363.
11. Kulkarni J, Wright S, Toole C, Morris J, Hirons R. Falls in patients with lower limb amputations: prevalence and contributing factors. *Physiotherapy*. 1996;82(2):130-136.
12. Gauthier-Gagnon C, Grise MC, Potvin D. Enabling factors related to prosthetic use by people with transtibial and transfemoral amputation. *Arch Phys Med Rehabil*. 1999;80(6):706-713.

1
2 461 13. Wong CK, Chen CC, Blackwell WM, Rahal RT, Benoy SA. Balance ability measured with
3 462 the Berg balance scale: a determinant of fall history in community-dwelling adults with leg
4 463 amputation. *J Rehabil Med*. 2015;47(1):80-86.
5
6 464 14. Miller WC, Speechley M, Deathe B. The prevalence and risk factors of falling and fear of
7 465 falling among lower extremity amputees. *Arch Phys Med Rehabil*. 2001;82(8):1031-1037.
8
9 466 15. Wong C, Chihuri S, Li G. Risk of fall-related injury in people with lower limb amputations:
10 467 A prospective cohort study. *J Rehabil Med*. 2016;48(1):80-85.
11
12 468 16. Mundell B, Maradit Kremers H, Visscher S, Hoppe K, Kaufman K. Direct medical costs of
13 469 accidental falls for adults with transfemoral amputations. *Prosthet Orthot Int*.
14 470 2017;41(6):564-570.
15
16 471 17. Chihuri S, Wong CK. Factors associated with the likelihood of fall-related injury among
17 472 people with lower limb loss. *Inj Prev*. 2018;5(1):1-8.
18
19 473 18. Roach KE. Measurement of health outcomes: Reliability, validity and responsiveness. *J*
20 474 *Prosthet Orthot*. 2006;18:8-12.
21
22 475 19. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Medicine*.
23 476 2000;30(1):1-15.
24
25 477 20. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement
26 478 studies (GRRAS). *J Clin Epidemiol*. 2011;64(1):96-106.
27
28 479 21. Balk EM, Gazula A, Markozannes G, et al. Lower limb prostheses: measurement
29 480 instruments, comparison of component effects by subgroups, and long-term outcomes.
30 481 AHRQ. Comparative effectiveness review number 213. 2018.
31
32 482 22. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*.
33 483 1979;86(2):420-428.
34
35 484 23. Hayen A, Dennis RJ, Finch CF. Determining the intra- and inter-observer reliability of
36 485 screening tools used in sports injury research. *J Sci Med Sport*. 2007;10(4):201-210.
37
38 486 24. Streiner DL, Norman GR, Cairney J. *Health measurement scales: a practical guide to their*
39 487 *development and use*. 5th ed. Oxford University Press. 2014.
40
41 488 25. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures
42 489 for use in clinical trials. *Health Technol Assess*. 1998;2(14):1-74.
43
44 490 26. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement
45 491 properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.
46
47 492 27. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for
48 493 patient-reported outcome measures used in patient-centered outcomes and comparative
49 494 effectiveness research. *Qual Life Res*. 2013;22(8):1889-1905.
50
51 495 28. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD. What is sufficient evidence for the
52 496 reliability and validity of patient-reported outcome measures? *Value Health*. 2007;10:S94-
53
54
55
56
57
58
59
60

S105.

29. Harvill LM. Standard error of measurement. *Educ Meas*. 1991;10(2):33-41.

30. Vincent WJ, Weir JP. *Statistics in Kinesiology*. Human Kinetics; 2012.

31. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol*. 1999;52(9):861-873.

32. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Phys Ther*. 1996;76(4):359-365.

33. Resnik L, Borgia M. Reliability of outcome measures for people with lower-limb amputations: Distinguishing true change from statistical error. *Phys Ther*. 2011;91(4):555-565.

34. Jayakaran P, Johnson GM, Sullivan SJ. Reliability and concurrent validity of the step quick turn test in older persons with a unilateral transtibial amputation. *Am J Phys Med Rehabil*. 2011;90(10):798-804.

35. Schoppen T, Boonstra A, Groothoff JW. The Timed “up and go” test: reliability and validity in persons with unilateral lower limb amputation. *Arch Phys Med Rehabil*. 1999;80(7):825-828.

36. Clemens SM, Gailey RS, Bennett CL, Pasquina PF, Kirk-Sanchez NJ, Gaunaud IA. The Component Timed-Up-and-Go test: the utility and psychometric properties of using a mobile application to determine prosthetic mobility in people with lower limb amputations. *Clin Rehabil*. 2018;32(3):388-397.

37. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19(4):539-549.

38. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231-240.

39. Sawers A, Hafner B. Validation of the Narrowing Beam Walking Test in lower limb prosthesis users. *Arch Phys Med Rehabil*. 2018;99(8):1491-1498.

40. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc*. 1991;39(2):142-148.

41. Dite W, Temple VA. A clinical test of stepping and change of direction to identify multiple falling older adults. *Arch Phys Med Rehabil*. 2002;83(11):1566-1571.

42. Bohannon RW. Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age and Ageing*. 1997;26(1):15-19.

43. Elm von E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational

1
2 533 Studies in Epidemiology (STROBE) statement: guidelines for reporting observational
3 534 studies. *Plos Med*. 2007;4(10):e296.
4
5 535 44. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data
6 536 capture (REDCap): A metadata-driven methodology and workflow process for providing
7 537 translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.
8
9 538 45. Doros G, Lew R. Design based on intra-class correlation coefficients. *Am J Biostat*. 2010.
10
11 539 46. Fewtrell MS, Kennedy K, Singhal A, et al. How much loss to follow-up is acceptable in
12 540 long-term randomized trials and prospective studies? *Arch Dis Child*. 2008;93(6):458-461.
14
15 541 47. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomized controlled trials.
16 542 *BMJ*. 2006;332(7547):969-971.
17
18 543 48. *Palmento Goverment Benefits Administrators. Lower limb prostheses. DMERC Medicare*
19 544 *Advis December 1994;(9):99-105*.
20
21 545 49. Hafner BJ, Gaunaud IA, Morgan SJ, Amtmann D, Salem R, Gailey RS. Construct validity
22 546 of the Prosthetic Limb Users Survey of Mobility (PLUS-M) in adults with lower limb
23 547 amputation. *Arch Phys Med Rehabil*. 2016;98(2):277-285.
25
26 548 50. Chaudhry S, Jin L, Meltzer D. Use of a self-report-generated Charlson Comorbidity Index
27 549 for predicting mortality. *Medical Care*. 2005;43(6):607-615.
28
29 550 51. Hanspal RS, Fisher K, Nieveen R. Prosthetic socket fit comfort score. *Disabil Rehabil*.
30 551 2003;25(22):1278-1280.
32
33 552 52. Hafner BJ, Morgan SJ, Askew RL, Salem R. Psychometric evaluation of self-report
34 553 outcome measures for prosthetic applications. *J Rehabil Res Dev*. 2016;53(6):797-812.
35
36 554 53. Dite W, Connor HJ, Curtis HC. Clinical identification of multiple fall risk early after
37 555 unilateral transtibial amputation. *Arch Phys Med Rehabil*. 2007;88(1):109-114.
38
39 556 54. Deathe AB, Miller WC. The L test of functional mobility: measurement properties of a
40 557 modified version of the timed “up & go” test designed for people with lower-limb
41 558 amputations. *Phys Ther*. 2005;85(7):626-635
43
44 559 55. Nanda R, Gupta S, Kanapathipillai P, Liow RYL, Rangan A. An assessment of the inter
45 560 examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity.
46 561 *Eur J Orthop Surg Traumatol*. 2008;18(7):495-500.
47
48 562 56. Defloor T, Schoonhoven L. Inter-rater reliability of the EPUAP pressure ulcer classification
49 563 system using photographs. *J Clin Nurs*. 2004;13(8):952-959.
50
51 564 57. Paulsen A, Overgaard S, Lauritsen JM. Quality of data entry using single entry, double
52 565 entry and automated forms processing: An example based on a study of patient-reported
53 566 outcomes. *PLoS ONE*. 2012;7(4):e35087.
55
56 567 58. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples).
57 568 *Biometrika*. 1965;52(3):591-611.
58
59
60

59. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther.* 1997;77(7):745-750.
60. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psycho Methods.* 1996;1(1):30-46.
61. Nunnally JC. *I. H. Bernstein. 1994. Psychometric Theory.*
62. Pezzin LE, Dillingham TR, MacKenzie EJ, Ephraim P, Rossbach P. Use and satisfaction with prosthetic limb devices and related services. *Arch Phys Med Rehabil.* 2004;85(5):723-729.
63. Ziegler-Graham K, MacKenzie EJ, Ephraim PL, Travison TG, Brookmeyer R. Estimating the prevalence of limb loss in the United States: 2005 to 2050. *Arch Phys Med Rehabil.* 2008;89(3):422-429.
64. Ephraim PL, Wegener ST, MacKenzie EJ, Dillingham TR, Pezzin LE. Phantom pain, residual limb pain, and back pain in amputees: Results of a national survey. *Arch Phys Med Rehabil.* 2005;86(10):1910-1919.
65. Wurdeman SR, Stevens PM, Campbell JH. Mobility Analysis of Amputees II: Comorbidities and mobility in lower limb prosthesis users. *Am J Phys Med Rehabil.* 2018;97(11):782-788.
66. Boonstra AM, Fidler V, Eisma WH. Walking speed of normal subjects and amputees: aspects of validity of gait analysis. *Prosthet Orthot Int.* 1993;17(2):78-82.
67. Moore M, Barker K. The validity and reliability of the four square step test in different adult populations: a systematic review. *Syst Rev.* 2017;6(1):1-9.
68. Beninato M, Portney LG. Applying concepts of responsiveness to patient management in neurologic physical therapy. *J Neurol Phys Ther.* 2011;35(2):75-81.
69. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155-163.
70. Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical change of significance: issues and methods. *Am J Occup Ther.* 1988;42(3):156-163.
71. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57(10):1008-1018.
72. Sawers A, Hafner BJ. Narrowing beam-walking is a clinically feasible approach for assessing balance ability in lower-limb prosthesis users. *J Rehabil Med.* 2018;50(5):457-464.
73. Sawers A, Hafner BJ. Conventional administration and scoring procedures suppress the diagnostic accuracy of a performance-based test designed to assess balance ability in lower limb prosthesis users. *Prosthet Orthot Int.* 2019;43(4):402-408.

1
2 605
3 606
4 607
5
6 608
7
8 609
9
10
11 610
12
13 611
14
15 612
16
17 613
18
19 614
20
21 615
22
23 616
24
25 617
26
27 618
28
29 619
30
31 620
32
33 621
34
35 622
36
37 623
38
39 624
40
41 625
42
43 626
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

74. Sawers A, Hafner BJ. Using clinical balance tests to assess fall risk among established unilateral lower limb prosthesis users: Cutoff scores and associated validity indices. *PM R.* 2019 (in press).

For Peer Review Only

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

Amputation Characteristics			Prosthetic Characteristics		Activity Characteristics		
Level	Etiology	Time Since Amputation (years)	Components <i>Feet</i> <i>Knees</i>	Time Since Delivery (years)	MFCL	PLUS-M (<i>T-score</i>)	Typical Prosthetic Use (<i>hrs/day</i>)
Participants (/60)	# of participants (/60)	Mean (SD)	Participants (60)	Mean (SD)	Participants (/60)	Mean (95% CI)	Mean (SD)
Transtibial (n=45) Transfemoral (n=15)	Trauma (n=26)	16.0 (14.8)	Energy Storing (53)	3.5 (3.0)	K2 (n=18) K3 (n=34) K4 (n=8)	56.1 (1.77)	13.9 (3.5)
	Infection (n=13)		Single-Axis (3)				
	Dysvascular (n=11)		Multi-Axis (2)				
	Tumor (n=9)		SACH (2)				
	Other (n=1)		Microprocessor (13)				
			Non-Microprocessor (2)				

CI: Confidence Interval, IQR: Interquartile range, MFCL: Medicare Functional Classification Level (K-level), SACH: solid ankle cushioned heel

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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

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

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

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

* 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

1
2 668 **Appendix 1**

3
4 669
5
6 670 **Timed Up and Go (TUG) Protocol**

8
9 671 Explanation to participant: The goal of this test is to rise from the chair, walk around the cone, walk
10
11 672 back to the chair and sit down again. You should walk at your normal, comfortable pace. I will time
12
13 673 you while you perform the test.

14
15 674
16
17
18 675 Demonstration: Demonstrate the test one time.

19
20 676
21
22 677 Test instructions: Begin the test sitting with your back against the back of the chair and your arms
23
24 678 resting on the armrests. When I say go, please stand up and walk around the cone, walk back to the
25
26 679 chair, and sit down again. Please walk at your normal, comfortable pace.

27
28 680
29
30
31 681 Practice: Administer one practice trial. Do not time the practice trial.

32
33 682
34
35
36 683 Administration: Administer the test 2 times. Begin timing when you say go. Stop timing when the
37
38 684 participants' buttocks touch the chair.

39
40 685
41
42
43 686 Scoring: Select faster of the two timed trials as the TUG score.

44
45 687

46
47 688

48
49 689

50
51 690

52
53 691

54
55 692

56
57 693

58
59 694

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.

1
2 718 **Narrowing Beam Walking Test (NBWT) Protocol**

3
4 719 Explanation to participant: The goal of this test is to walk as far as possible along the beam. Speed is
5
6 720 not being evaluated. Begin the test by standing with one foot on the wide end of the beam and the
7
8
9 721 other foot on the ground to the side. You may choose which foot to put on the beam and which to put
10
11 722 on the ground. Please cross both your arms across your chest.

12
13 723
14
15 724 Demonstration: Demonstrate the test one time.

16
17
18 725
19
20 726 Test instructions: When I say go, please walk along the beam as far as you can. Please walk at a
21
22 727 comfortable speed. Remember to keep your arms crossed over your chest as you walk. Once you
23
24
25 728 move your arms away from your body or step off the beam, I will ask you to stop.

26
27 729
28
29 730 Practice: Do not administer a practice trial.

30
31
32 731
33
34 732 Administration: Administer the test 5 times. Stop the trial when a participant: walks the length of the
35
36 733 beam, steps off the beam, or moves their arms away from their body.

37
38
39 734
40
41 735 Scoring: Average the distances walked during trials 3 through 5, and divide by 22. This normalized
42
43 736 distance (i.e., 0-1) is recorded as the NBWT score. The average distance is divided by 22 and not 24
44
45 737 (the total length of the beam), because participants begin the test with one foot on the beam and must
46
47
48 738 walk past the 0.61m (2.0ft) mark to receive a score above 0.0.

49
50 739
51
52 740
53
54
55 741
56
57 742
58
59
60

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.

1
2 1 **Running Head:** Reliability of balance tests in amputees

3
4 2
5
6 3 **Title:** Inter-rater and test-retest reliability of performance-based clinical balance tests administered to
7
8 4 established unilateral lower limb prosthesis users

9
10
11 5
12
13 6 **Authors** Andrew Sawers, CPO, PhD¹, Janis Kim, MPT¹, Geoff Balkman², Brian Hafner, PhD²

14
15 7
16
17 8 **Author Contribution:**

18
19
20 9 *Study Concept and Design:* A. Sawers, B. Hafner

21
22 10 *Acquisition of Data:* A. Sawers, J. Kim, G. Balkman, B. Hafner

23
24 11 *Analysis and Interpretation of Data:* A. Sawers, B. Hafner

25
26 12 *Drafting of Manuscript:* A. Sawers

27
28 13 *Critical Revision of Manuscript of Important Intellectual Content:* B. Hafner, J. Kim, G. Balkman

29
30 14 *Statistical Analysis:* A. Sawers

31
32 15 *Study Supervision:* A. Sawers

33
34 16
35
36 17 **Institutions and Affiliations:**

37
38
39 18 ¹ Department of Kinesiology, University of Illinois at Chicago, Chicago, IL 60612

40
41 19 ² Department of Rehabilitation Medicine, University of Washington, Seattle, WA 98195

42
43 20
44
45 21
46
47 22
48
49 23 **Prior Publication:**

50
51
52 24 The work described in this manuscript has not been previously published, nor is it under consideration
53
54
55 25 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 (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., $\geq .98$). 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., $ICC \geq .70$ and $ICC \geq .90$, 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

1
2 101 **Introduction**

3
4 102 The use of valid and reliable instruments to document clinical outcomes is increasingly expected **by**
5
6 103 **employers and third-party payers** in order to justify and evaluate rehabilitation services and
7
8 104 interventions provided to lower limb prosthesis (LLP) users¹⁻³. Despite the demand for robust
9
10 105 documentation, several barriers inhibit regular use of standardized outcome measures in clinical
11
12 106 practice²⁻⁶. Selection of suitable outcome measures for different clinical applications^{5,7-9}, and
13
14 107 interpretation of test scores to guide clinical decision-making^{6,7} are frequently cited as challenges to
15
16 108 routine use of outcome measures in clinical practice. These barriers may limit, among other things,
17
18 109 regular assessment of balance ability and fall risk among LLP users, potentially contributing to the
19
20 110 persistence of falls¹⁰⁻¹⁴, and their related consequences¹⁵⁻¹⁷ among LLP users.
21
22
23
24

25 111
26
27 112 Indices of reproducibility (e.g., estimates of reliability, error, and detectable change), in the
28
29 113 population of interest (i.e., LLP users)¹⁸⁻²⁰, are needed to formulate recommendations for performance-
30
31 114 based clinical ~~balance~~ tests suited to the care of LLP users^{18,19,21}. Inter-rater and test-retest reliability,
32
33 115 often quantified using intraclass correlation coefficients (ICC)²², are estimates of the stability of
34
35 116 measurement between raters and over a period of time when no changes in health are believed to have
36
37 117 occurred^{19,23,24}. ICCs therefore provide test administrators with important information about the
38
39 118 consistency of measurement(s) produced by a test within a given population. The magnitude of
40
41 119 reliability estimates can also be used to guide clinicians in selecting performance-based tests for specific
42
43 120 applications. To be recommended for group-level comparisons, a clinical test should generally
44
45 121 demonstrate a test-retest ICC of .70 or greater²⁵⁻²⁸. For applications that involve individual-level
46
47 122 decisions, a test should have higher reliability, recommended as .90 or greater^{19,25,28}. Reliability
48
49 123 estimates do not however help clinicians interpret test scores or provide the information needed to make
50
51 124 evidence-based decisions for individual patients. The interpretation of individual test scores requires
52
53 125 knowledge of the standard error of measurement (SEM) and minimal detectable change (MDC) of a test.
54
55
56
57
58
59
60

The SEM reflects the error associated with a *single* measurement¹⁹. 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 individual³¹. 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 test³². 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 reproducibility³³⁻³⁶. **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 small³³⁻³⁵, and often below recommendations for establishing reliability indices³⁷. Second, evidence of reproducibility for some tests is dated and may not reflect best practices for assessing test-retest reliability³⁵. 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**

1
2 151 **and its components are suitable for clinical practice.** Finally, given the relative nature of that test-
3
4 152 retest reliability **is relative to the sample of participants from which it is derived**³⁸, and the
5
6 153 **evaluation of a test's reliability is a continual process (i.e., tests are not reliable or unreliable, they**
7
8
9 154 **exhibit more or less evidence of reliability)**, it is beneficial to **accumulate additional evidence of**
10
11 155 **reliability for a test, by comparing, contrasting, and confirming values across different samples of LLP**
12
13 156 users. There is therefore insufficient psychometric evidence to generate recommendations for the
14
15 157 selection and interpretation of common performance-based clinical ~~balance~~ tests administered to LLP
16
17
18 158 users.
19

20 159
21
22 160 The objective of this study was therefore to assess the reproducibility of performance-based clinical
23
24
25 161 ~~balance~~ tests administered to established unilateral LLP users, deriving estimates of reliability indices to
26
27 162 help clinicians and researchers select and interpret the studied tests. Specifically, the ICCs for inter-rater
28
29 163 and test-retest reliability were assessed, from which the standard SEM and MDC of each of the studied
30
31
32 164 tests were derived. We hypothesized that the studied tests (i.e., NBWT, FSST, TUG, and 10mWT)
33
34 165 would demonstrate evidence of reliability matching or exceeding that required for group-level
35
36 166 comparisons (i.e., $ICC \geq 0.7$) and individual decision-making (i.e., $ICC \geq 0.9$) in assessments of balance
37
38
39 167 and fall risk.
40

41 168
42
43 169 **Methods**
44

45 170 **Study Design**
46

47
48 171 A **two-site (Chicago and Seattle)** repeated measures study was conducted between February 2018 and
49
50 172 November 2018 to assess the reproducibility of the Narrowing Beam Walking Test (NBWT)³⁹, Timed
51
52 173 Up and Go (TUG)⁴⁰, Four Square Step Test (FSST)⁴¹, and the 10 meter Walk Test (10mWT)⁴² when
53
54
55 174 administered to LLP users. Best practices in conducting observational studies (i.e., STROBE:
56
57 175 Strengthening The Reporting of Observational studies in Epidemiology⁴³), as well as the collection and
58
59
60

reporting of reliability data (i.e., CONsensus-based Standards for the selection of health Measurement INstruments (COSMIN))^{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 Chicago⁴⁴. 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 .80⁴⁵. Psychometric standards, however, recommend a minimum sample of 50 participants to estimate reliability statistics¹⁹. Based on the more conservative estimate (i.e., recommended psychometric standards), and planning for modest attrition (i.e., 20%) between test and retest sessions^{46,47}, a sample of 60 LLP users was targeted.

Procedures

1
2 201 **At their first visit**, all participants were administered demographic and characterization measures,
3
4 202 as well as four performance-based clinical ~~balance~~ tests. **At the end of the first test session**, each
5
6 203 participant was scheduled a retest session three to nine days **later**. This test-retest interval was chosen to
7
8 204 provide sufficient time to limit familiarity with the tests, and to avoid the possibility of changes in health
9
10
11 205 or prosthetic fit. Participants wore their desired prosthesis and footwear, both of which were held
12
13 206 constant between test sessions.

14
15 207
16
17
18 208 **Measurements**

19
20 209 *Participant demographic and characterization measures*

21
22 210 Sociodemographic information including age, height, weight, gender, race, and ethnicity was
23
24
25 211 collected via self-report. Amputation characteristics (i.e., level, etiology, and time since amputation),
26
27 212 prosthetic-related factors (i.e., components, and time since delivery), and mobility (i.e., Medicare
28
29 213 Functional Classification Level (MFCL)⁴⁸, Prosthetic Limb Users Survey of Mobility (PLUS-M)⁴⁹, and
30
31 214 typical prosthesis use/day) were collected via self-report, interview, and/or clinical inspection. The
32
33
34 215 number of comorbid conditions was collected via self-report using the Charlson Comorbidity Index
35
36 216 (CCI)⁵⁰. Comfort of the prosthesis worn by each participant was evaluated using the Socket Comfort
37
38
39 217 Score (SCS), **scored from 0 to 10**^{51,52}. At retest, participants were asked whether they had experienced
40
41 218 changes in their prosthesis and/or health since the first testing session. The SCS was re-administered at
42
43 219 retest to confirm that participants' scores were within the MDC₉₀ of the SCS in LLP users (i.e., ± 2.82
44
45 220 points)⁵².

46
47
48 221
49
50 222 *Administration of performance-based clinical ~~balance~~ tests*

51
52 223 The TUG⁴⁰, FSST⁴¹ 10mWT⁴², and the NBWT³⁹ were administered **in a randomized order**, and
53
54
55 224 scored according to standardized instructions (Appendix 1). **The same test sequence was used in both**
56
57 225 **test sessions**. Each balance test was selected because it could be performed in less than five minutes, has
58
59
60

been designed or used to assess fall risk in other populations, and has evidence of validity among LLP users^{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 sites⁵⁵. Further, no **communication was allowed between concurrent raters during testing to limit** their awareness of each other's assessments, which could influence scoring⁵⁶. Finally, data collected on standardized paper forms were double entered by research staff to minimize data-entry errors⁵⁷. **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 indices²⁰. **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 test⁵⁸. Measures of central tendency and dispersion were calculated for each variable based on their

1
2 251 distribution. Paired 2-sided t-tests were used to test for differences in test scores between raters, and
3
4 252 between test sessions (i.e., effect of time). A Bonferroni correction ($\alpha=.0125$) was used for multiple
5
6 253 comparisons within the sample.
7

8
9 254
10
11 255 Inter-rater reliability of each clinical ~~balance~~ test was evaluated by calculating the ICC (1,1) from
12
13 256 measurements taken simultaneously by two raters per participant^{22,38,59}. Each clinical ~~balance~~ test was
14
15 257 also evaluated for **absolute** test-retest reliability by calculating the ICC (2,1) from measurements taken
16
17
18 258 on two visits by the same rater^{22,38}. 95% confidence intervals (CI) were derived for each ICC using the
19
20 259 F-distribution⁶⁰. Test-retest ICCs were compared to established thresholds in order to determine the
21
22 260 suitability of each test for clinical and research applications (i.e., .70 or greater for group-level
23
24
25 261 comparisons²⁵⁻²⁸, and .90 or greater for individual-level decisions^{19,25-28}).
26
27 262
28

29 263 Standard error of measurement (SEM) estimates were calculated as $SEM = SD * \sqrt{(1 - ICC)}$,
30
31
32 264 where SD is the pooled standard deviation of test scores across both test sessions from the same rater
33
34 265 (i.e., rater A)³⁰, and ICC is the test-retest intraclass correlation coefficient^{29,33}. Alternatives sources of
35
36 266 the standard deviation were explored (e.g., standard deviation from the first test session³³, and total sum
37
38
39 267 of the squares from a one-way ANOVA^{33,38,59}), but neither resulted in an appreciable change in SEM
40
41 268 values. MDC estimates were calculated as $MDC_{90} = 1.64 * \sqrt{SEM} * \sqrt{2}$ where 1.64 is the z-score for a
42
43 269 90% confidence interval³⁸.
44
45
46 270
47

48 271 **Results**
49

50 272 Sixty ambulatory unilateral LLP users with mean age of 55.1 years (SD = 21.3 years), 42 of which
51
52 273 were male, were recruited and completed the study. Amputation, prosthesis, and mobility characteristics
53
54
55 274 of the study sample are presented in Table 1. 48% of participants reported no co-morbid conditions,
56
57 275 while 20%, 13%, and 10% reported one, two, or three or more co-morbid conditions (e.g., asthma,
58
59
60

diabetes, HIV/AIDS), respectively. Observed test scores (Supplement 1) for the FSST, TUG, and 10mWT were non-normally distributed (Shapiro-Wilk $W=.798-.919$, $p\leq.001$). In contrast, NBWT scores were normally distributed (Shapiro-Wilk $W=.957-.974$, $p\geq.195$). 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., $n>30$)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., $\geq.98$) 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 MDC₉₀ 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 MDC₉₀ for the NBWT, a test of normalized walking distance, were .07 and .16, respectively.

1
2 301
3
4 302
5
6 303
7
8
9 304
10
11 305
12
13 306
14
15 307
16
17
18 308
19
20 309
21
22 310
23
24
25 311
26
27 312
28
29 313
30
31
32 314
33
34 315
35
36 316
37
38
39 317
40
41 318
42
43 319
44
45 320
46
47
48 321
49
50 322
51
52 323
53
54
55 324
56
57 325
58
59
60

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., $ICC \geq .70$), while the FSST, 10mWT, and NBWT are also suitable for individual-level applications among established unilateral LLP users (i.e., $ICC \geq .90$). **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., $ICC \geq 0.7$) and individual decision-making (i.e., $ICC \geq 0.9$).** 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., MDC_{90}). 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 ICCs³⁸. 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, .55³⁶, and smaller than another, 1.6³³. 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), respectively⁶⁸. To date, the SEM for the 10mWT or FSST among LLP users has not been reported.

Point estimates for the MDC₉₀ 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 MDC₉₀ point estimates in this study and those reported

1
2 351 previously, the 95% CIs of MDC₉₀ derived here overlapped with previously reported values. This
3
4 352 implies that while diverse, there is some consistency with regard to what constitutes a true change in
5
6 353 performance over time within each of the studied tests.
7

8
9 354
10
11 355 This study provides the first evidence of inter-rater reliability using ICCs for the TUG, FSST, and
12
13 356 10mWT in LLP users (Table 2). The reported inter-rater ICC of the TUG (i.e., .99 [95% CI: .99-1.0]) is
14
15 357 consistent with what has been reported using a different approach (i.e., .96 with Pearson's r)³⁵.
16
17
18 358

19
20 359 **Application and interpretation of reliability indices**
21

22 360 The reliability indices reported here can serve to assist clinicians and researchers in selecting and
23
24
25 361 interpreting performance-based clinical ~~balance~~ tests³⁸. ICCs provide an indication of how confident
26
27 362 users should be in their measurements, namely, what percentage of the observed score variance is due to
28
29 363 true score variance versus that attributable to error. All of the studied tests were found to have inter-rater
30
31
32 364 and test-retest ICCs between .98-.99 and .88-.97 and respectively (Tables 2, 3). This suggests that 98%
33
34 365 to 99% of the observed score variance between raters, and 88% to 97% of observed score variance
35
36 366 between test sessions is due to true score variance, with the remaining observed score variance
37
38
39 367 attributable to error (i.e., 1% to 12%). Trained users familiar with these tests can therefore be confident
40
41 368 in the individual measurements made with these performance-based tests. Beyond establishing a level of
42
43 369 confidence in an instrument, the magnitude of a test-retest ICC also suggests whether a test may be
44
45 370 suitable for group-level comparisons (i.e., $ICC \geq .70$)²⁵⁻²⁸, and/or individual-level applications (i.e., ICC
46
47
48 371 $\geq .90$)^{19,25,28}. With the exception of the TUG, studied tests (i.e., FSST, NBWT, and 10mWT) had test-
49
50
51 372 retest ICCs $\geq .90$, indicating that they are each suitable for group as well as individual-level applications.
52
53 373 Because the TUG test-retest ICC (.88) is below, albeit slightly, the recommended threshold of 0.9, users
54
55 374 should be aware that this test has greater error relative to the other performance tests studied here, **and**
56
57 375 **may therefore be better suited to group-level comparisons rather than making decisions about**
58
59
60

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 coefficient⁶⁹, determines if reliability results can be generalized beyond the participants and raters included in the study³⁸. Consistent with most reproducibility studies conducted to-date^{33,34,36}, test-retest ICCs in the present study were calculated using ICC (2,1), which includes both random and systematic error^{22,38}. Model 2 permits the generalization of reliability results to other raters with similar participants or patients³⁸. 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^{29,38}. 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)^{70,71} are above and beyond changes that may be attributable to measurement error³¹. 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 MDC₉₀ of .16, indicating that a real change in balance ability had been observed.

1
2 401
3
4 402
5
6 403
7
8
9 404
10
11 405
12
13 406
14
15 407
16
17
18 408
19
20 409
21
22 410
23
24
25 411
26
27 412
28
29 413
30
31
32 414
33
34 415
35
36 416
37
38 417
39
40
41 418
42
43 419
44
45 420
46
47
48 421
49
50 422
51
52 423
53
54
55 424
56
57 425
58
59
60

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^{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)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 performance⁷³. It is possible that changes to TUG administration and scoring procedures may be able to resolve these issues⁷³. 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 error³⁸. 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 $\geq$.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 $\geq$.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

1. Heinemann AW, Connelly L, Ehrlich-Jones L, Fatone S. Outcome instruments for prosthetics: clinical applications. *Phys Med Rehabil Clin N Am*. 2014;25(1):179-198.
2. Wedge FM, Braswell-Christy J, Brown CJ, Foley KT, Graham C, Shaw S. Factors influencing the use of outcome measures in physical therapy practice. *Physiother Theory Pract*. 2012;28(2):119-133.
3. Jette DU, Halbert J, Iverson C, Miceli E. Use of standardized outcome measures in physical therapist practice: perceptions and applications. *Phys Ther*. 2009;89(2):125-135.
4. Bowman J. Challenges to measuring outcomes in occupational therapy: A qualitative focus group study. *Br J Occup Ther*. 2006;69(10):464-472.
5. Duncan EAS, Murray J. The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Serv Res*. 2012;12(1):96-104.
6. Gaunaud I, Spaulding SE, Amtmann D, et al. Use of and confidence in administering outcome measures among clinical prosthetists: Results from a national survey and mixed-methods training program. *Prosthet Orthot Int*. 2015;39(4):314-321.
7. Hafner BJ, Spaulding SE, Salem R, Morgan SJ, Gaunaud I, Gailey R. Prosthetists' perceptions and use of outcome measures in clinical practice: Long-term effects of focused continuing education. *Prosthet Orthot Int*. 2016;41(3):1-8.
8. Stapleton T, McBrearty C. Use of Standardized assessments and outcome measures among a sample of Irish occupational therapists working with adults with physical disabilities. *Br J Occup Ther*. 2009;72(2):55-64.
9. Belazi D, Goldfarb NI, He H. Measuring health-related quality of life in the clinical setting. *Expert Rev Pharmacoecon Outcomes Res*. 2014;2(2):109-117.
10. Kim J, Major MJ, Hafner B, Sawers A. Frequency and circumstances of falls reported by ambulatory unilateral lower limb prosthesis users: A secondary analysis. *PM R*. 2019;11(4):344-363.
11. Kulkarni J, Wright S, Toole C, Morris J, Hirons R. Falls in patients with lower limb amputations: prevalence and contributing factors. *Physiotherapy*. 1996;82(2):130-136.
12. Gauthier-Gagnon C, Grise MC, Potvin D. Enabling factors related to prosthetic use by people with transtibial and transfemoral amputation. *Arch Phys Med Rehabil*. 1999;80(6):706-713.

13. Wong CK, Chen CC, Blackwell WM, Rahal RT, Benoy SA. Balance ability measured with the Berg balance scale: a determinant of fall history in community-dwelling adults with leg amputation. *J Rehabil Med*. 2015;47(1):80-86.
14. Miller WC, Speechley M, Deathe B. The prevalence and risk factors of falling and fear of falling among lower extremity amputees. *Arch Phys Med Rehabil*. 2001;82(8):1031-1037.
15. Wong C, Chihuri S, Li G. Risk of fall-related injury in people with lower limb amputations: A prospective cohort study. *J Rehabil Med*. 2016;48(1):80-85.
16. Mundell B, Maradit Kremers H, Visscher S, Hoppe K, Kaufman K. Direct medical costs of accidental falls for adults with transfemoral amputations. *Prosthet Orthot Int*. 2017;41(6):564-570.
17. Chihuri S, Wong CK. Factors associated with the likelihood of fall-related injury among people with lower limb loss. *Inj Prev*. 2018;5(1):1-8.
18. Roach KE. Measurement of health outcomes: Reliability, validity and responsiveness. *J Prosthet Orthot*. 2006;18:8-12.
19. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Medicine*. 2000;30(1):1-15.
20. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS). *J Clin Epidemiol*. 2011;64(1):96-106.
21. Balk EM, Gazula A, Markozannes G, et al. Lower limb prostheses: measurement instruments, comparison of component effects by subgroups, and long-term outcomes. AHRQ. Comparative effectiveness review number 213. 2018.
22. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.
23. Hayen A, Dennis RJ, Finch CF. Determining the intra- and inter-observer reliability of screening tools used in sports injury research. *J Sci Med Sport*. 2007;10(4):201-210.
24. Streiner DL, Norman GR, Cairney J. *Health measurement scales: a practical guide to their development and use*. 5th ed. Oxford University Press. 2014.
25. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess*. 1998;2(14):1-74.
26. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.
27. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res*. 2013;22(8):1889-1905.
28. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health*. 2007;10:S94-

- 1
2 523 S105.
3
- 4 524 29. Harvill LM. Standard error of measurement. *Educ Meas.* 1991;10(2):33-41.
5
- 6 525 30. Vincent WJ, Weir JP. *Statistics in Kinesiology*. Human Kinetics; 2012.
7
- 8 526 31. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based
9 527 criterion for identifying meaningful intra-individual changes in health-related quality of life.
10 528 *J Clin Epidemiol.* 1999;52(9):861-873.
11
- 12
13 529 32. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum
14 530 level of detectable change for the Roland-Morris questionnaire. *Phys Ther.* 1996;76(4):359-
15 531 365.
16
- 17 532 33. Resnik L, Borgia M. Reliability of outcome measures for people with lower-limb
18 533 amputations: Distinguishing true change from statistical error. *Phys Ther.* 2011;91(4):555-
19 534 565.
20
- 21
22 535 34. Jayakaran P, Johnson GM, Sullivan SJ. Reliability and concurrent validity of the step quick
23 536 turn test in older persons with a unilateral transtibial amputation. *Am J Phys Med Rehabil.*
24 537 2011;90(10):798-804.
25
- 26 538 35. Schoppen T, Boonstra A, Groothoff JW. The Timed “up and go” test: reliability and
27 539 validity in persons with unilateral lower limb amputation. *Arch Phys Med Rehabil.*
28 540 1999;80(7):825-828.
29
- 30
31 541 36. Clemens SM, Gailey RS, Bennett CL, Pasquina PF, Kirk-Sanchez NJ, Gaunaud IA. The
32 542 Component Timed-Up-and-Go test: the utility and psychometric properties of using a
33 543 mobile application to determine prosthetic mobility in people with lower limb amputations.
34 544 *Clin Rehabil.* 2018;32(3):388-397.
35
- 36 545 37. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the
37 546 methodological quality of studies on measurement properties of health status measurement
38 547 instruments: an international Delphi study. *Qual Life Res.* 2010;19(4):539-549.
39
- 40
41 548 38. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the
42 549 SEM. *J Strength Cond Res.* 2005;19(1):231-240.
43
- 44 550 39. Sawers A, Hafner B. Validation of the Narrowing Beam Walking Test in lower limb
45 551 prosthesis users. *Arch Phys Med Rehabil.* 2018;99(8):1491-1498.
46
- 47 552 40. Podsiadlo D, Richardson S. The timed “Up & Go”: a test of basic functional mobility for
48 553 frail elderly persons. *J Am Geriatr Soc.* 1991;39(2):142-148.
49
- 50
51 554 41. Dite W, Temple VA. A clinical test of stepping and change of direction to identify multiple
52 555 falling older adults. *Arch Phys Med Rehabil.* 2002;83(11):1566-1571.
53
- 54 556 42. Bohannon RW. Comfortable and maximum walking speed of adults aged 20-79 years:
55 557 reference values and determinants. *Age and Ageing.* 1997;26(1):15-19.
56
- 57 558 43. Elm von E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational
58
59
60

- Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Plos Med*. 2007;4(10):e296.
44. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.
 45. Doros G, Lew R. Design based on intra-class correlation coefficients. *Am J Biostat*. 2010.
 46. Fewtrell MS, Kennedy K, Singhal A, et al. How much loss to follow-up is acceptable in long-term randomized trials and prospective studies? *Arch Dis Child*. 2008;93(6):458-461.
 47. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomized controlled trials. *BMJ*. 2006;332(7547):969-971.
 48. *Palmento Goverment Benefits Administrators. Lower limb prostheses. DMERC Medicare Advis December 1994;(9):99-105.*
 49. Hafner BJ, Gaunaud IA, Morgan SJ, Amtmann D, Salem R, Gailey RS. Construct validity of the Prosthetic Limb Users Survey of Mobility (PLUS-M) in adults with lower limb amputation. *Arch Phys Med Rehabil*. 2016;98(2):277-285.
 50. Chaudhry S, Jin L, Meltzer D. Use of a self-report-generated Charlson Comorbidity Index for predicting mortality. *Medical Care*. 2005;43(6):607-615.
 51. Hanspal RS, Fisher K, Nieveen R. Prosthetic socket fit comfort score. *Disabil Rehabil*. 2003;25(22):1278-1280.
 52. Hafner BJ, Morgan SJ, Askew RL, Salem R. Psychometric evaluation of self-report outcome measures for prosthetic applications. *J Rehabil Res Dev*. 2016;53(6):797-812.
 53. Dite W, Connor HJ, Curtis HC. Clinical identification of multiple fall risk early after unilateral transtibial amputation. *Arch Phys Med Rehabil*. 2007;88(1):109-114.
 54. Deathe AB, Miller WC. The L test of functional mobility: measurement properties of a modified version of the timed "up & go" test designed for people with lower-limb amputations. *Phys Ther*. 2005;85(7):626-635
 55. Nanda R, Gupta S, Kanapathipillai P, Liow RYL, Rangan A. An assessment of the inter examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. *Eur J Orthop Surg Traumatol*. 2008;18(7):495-500.
 56. Defloor T, Schoonhoven L. Inter-rater reliability of the EPUAP pressure ulcer classification system using photographs. *J Clin Nurs*. 2004;13(8):952-959.
 57. Paulsen A, Overgaard S, Lauritsen JM. Quality of data entry using single entry, double entry and automated forms processing: An example based on a study of patient-reported outcomes. *PLoS ONE*. 2012;7(4):e35087.
 58. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3):591-611.

1
2 595 59. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an
3 596 applied example using elbow flexor strength data. *Phys Ther.* 1997;77(7):745-750.
4
5 597 60. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients.
6 598 *Psycho Methods.* 1996;1(1):30-46.
7
8 599 61. Nunnally JC. *I. H. Bernstein. 1994. Psychometric Theory.*
9
10 600 62. Pezzin LE, Dillingham TR, MacKenzie EJ, Ephraim P, Rossbach P. Use and satisfaction
11 601 with prosthetic limb devices and related services. *Arch Phys Med Rehabil.* 2004;85(5):723-
12 602 729.
13
14
15 603 63. Ziegler-Graham K, MacKenzie EJ, Ephraim PL, Travison TG, Brookmeyer R. Estimating
16 604 the prevalence of limb loss in the United States: 2005 to 2050. *Arch Phys Med Rehabil.*
17 605 2008;89(3):422-429.
18
19 606 64. Ephraim PL, Wegener ST, MacKenzie EJ, Dillingham TR, Pezzin LE. Phantom pain,
20 607 residual limb pain, and back pain in amputees: Results of a national survey. *Arch Phys Med*
21 608 *Rehabil.* 2005;86(10):1910-1919.
22
23
24 609 65. Wurdeman SR, Stevens PM, Campbell JH. Mobility Analysis of Amputees II:
25 610 Comorbidities and mobility in lower limb prosthesis users. *Am J Phys Med Rehabil.*
26 611 2018;97(11):782-788.
27
28 612 66. Boonstra AM, Fidler V, Eisma WH. Walking speed of normal subjects and amputees:
29 613 aspects of validity of gait analysis. *Prosthet Orthot Int.* 1993;17(2):78-82.
30
31
32 614 67. Moore M, Barker K. The validity and reliability of the four square step test in different adult
33 615 populations: a systematic review. *Syst Rev.* 2017;6(1):1-9.
34
35 616 68. Beninato M, Portney LG. Applying concepts of responsiveness to patient management in
36 617 neurologic physical therapy. *J Neurol Phys Ther.* 2011;35(2):75-81.
37
38 618 69. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients
39 619 for reliability research. *J Chiropr Med.* 2016;15(2):155-163.
40
41 620 70. Ottenbacher KJ, Johnson MB, Hojem M. The significance of clinical change and clinical
42 621 change of significance: issues and methods. *Am J Occup Ther.* 1988;42(3):156-163.
43
44
45 622 71. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID)
46 623 proportions facilitated group responsiveness comparisons using individual threshold
47 624 criteria. *J Clin Epidemiol.* 2004;57(10):1008-1018.
48
49 625 72. Sawers A, Hafner BJ. Narrowing beam-walking is a clinically feasible approach for
50 626 assessing balance ability in lower-limb prosthesis users. *J Rehabil Med.* 2018;50(5):457-
51 627 464.
52
53
54 628 73. Sawers A, Hafner BJ. Conventional administration and scoring procedures suppress the
55 629 diagnostic accuracy of a performance-based test designed to assess balance ability in lower
56 630 limb prosthesis users. *Prosthet Orthot Int.* 2019;43(4):402-408.
57
58
59
60

74. Sawers A, Hafner BJ. Using clinical balance tests to assess fall risk among established unilateral lower limb prosthesis users: Cutoff scores and associated validity indices. *PM R*. 2019 (in press).

For Peer Review Only

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

Amputation Characteristics			Prosthetic Characteristics		Activity Characteristics		
Level	Etiology	Time Since Amputation (years)	Components <i>Feet</i> <i>Knees</i>	Time Since Delivery (years)	MFCL	PLUS-M (T-score)	Typical Prosthetic Use (hrs/day)
Participants (/60)	# of participants (/60)	Mean (SD)	Participants (60)	Mean (SD)	Participants (/60)	Mean (95% CI)	Mean (SD)
Transtibial (n=45) Transfemoral (n=15)	Trauma (n=26)	16.0 (14.8)	Energy Storing (53)	3.5 (3.0)	K2 (n=18) K3 (n=34) K4 (n=8)	56.1 (1.77)	13.9 (3.5)
	Infection (n=13)		Single-Axis (3)				
	Dysvascular (n=11)		Multi-Axis (2)				
	Tumor (n=9)		SACH (2)				
	Other (n=1)		Microprocessor (13)				
			Non-Microprocessor (2)				

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

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

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

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

* 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.

1
2 719 **Four Square Step Test (FSST) Protocol**

3
4 720 Explanation to participant: The goal of this test is to step over the canes in a specific sequence as
5
6 721 quickly as possible. I will time you while you perform this test.

7
8
9 722
10
11 723 Demonstration: Demonstrate the test one time. Demonstrate starting in square 1 and stepping in
12
13 724 squares 2, 3, 4, 1, 4, 3, 2, and 1.

14
15 725
16
17
18 726 Test instructions: When I say go, please step in the sequent I demonstrated. Try to complete the
19
20 727 sequence as fast as possible without touching the canes. Both feet must make contact with the floor
21
22 728 in each square. Face forward during the entire sequence.

23
24
25 729
26
27 730 Practice: Administer one practice trial. Do not time the practice trial.

28
29 731
30
31
32 732 Administration: Administer the test 2 times. Begin timing when the first foot contacts square 2. Stop
33
34 733 timing when the last foot contacts square 1. Repeat the trial if the participant does not complete the
35
36 734 sequence, loses balance, or contact a cane.

37
38
39 735
40
41 736 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.

1
2 769 **10 Meter Walk Test (10MWT) Protocol**

3
4 770
5
6 771 Explanation to participant: The goal of this test is to walk a short distance at your preferred
7
8
9 772 comfortable walking speed. I will time you while you perform this test.

10
11 773
12
13 774 Demonstration: Demonstrate the test one time.

14
15
16 775
17
18 776 Test instructions: When I say ‘go,’ please walk at your normal, comfortable pace until I say stop.

19
20 777
21
22 778 Practice: Administer one practice trial.

23
24
25 779
26
27 780 Administration: Prepare a 14-meter walkway in a hallway or other unobstructed area. Place lines at
28
29
30 781 0, 2, 12, and 14 meters. Have the participant start in a standing position on the 0-meter line. Inform
31
32 782 them that, on the word “go,” they are to walk at a comfortable speed until you say “stop.” Begin
33
34 783 timing when the participant crosses the 2-meter line. Stop timing when the participant crosses the
35
36 784 12-meter line. Inform the participant to stop when they cross the 14-meter line. Repeat the test two
37
38
39 785 times.

40
41 786
42
43 787 Scoring: Select the faster of the two timed trials as the TUG score.

44
45 788

46
47 789

48
49 790

50
51

52
53

54
55

56
57

58
59
60

ID	Session 1					
	NBWT score		TUG score		FSST score	
	RaterA	RaterB	RaterA	RaterB	RaterA	RaterB
S1	0.60	0.60	8.36	7.94	4.78	4.89
S2	0.66	0.65	7.79	7.62	6.71	6.77
S3	0.83	0.84	9.20	9.24	6.94	7.12
S4	0.29	0.27	19.16	18.45	21.35	20.63
S5	0.78	0.78	7.77	7.63	6.82	6.96
S6	0.36	0.34	9.19	9.22	9.19	9.22
S7	0.29	0.30	14.02	14.12	10.94	11.03
S8	0.67	0.66	7.79	7.56	4.49	4.62
S9	0.76	0.76	7.40	7.77	6.49	6.36
S10	0.39	0.38	11.52	11.33	10.78	10.94
S11	0.27	0.26	13.42	13.44	12.96	12.45
S12	0.48	0.48	11.37	11.08	8.36	9.72
S13	0.81	0.81	8.19	8.02	6.41	6.26
S14	0.48	0.47	7.02	6.76	7.83	7.96
S15	0.30	0.27	14.82	14.96	13.90	14.07
S16	0.20	0.16	14.68	14.75	13.02	12.78
S17	0.33	0.33	9.65	9.64	9.03	9.22
S18	0.27	0.26	12.08	12.26	13.30	13.21
S19	0.34	0.30	10.63	10.59	10.06	10.15
S20	0.36	0.36	8.96	8.71	7.58	7.58
S21	0.48	0.48	10.36	10.35	9.02	9.02
S22	0.52	0.54	11.44	11.28	9.19	9.09
S23	0.41	0.40	10.67	10.51	11.69	11.78
S24	0.45	0.46	8.78	8.84	9.57	9.69
S25	0.14	0.13	20.58	20.72	32.89	32.94
S26	0.30	0.06	14.34	14.25	13.78	13.63
S27	0.25	0.25	8.65	8.50	8.39	8.35
S28	0.40	0.39	13.91	14.50	20.83	21.00
S29	0.18	0.18	9.96	10.01	17.13	17.21
S30	0.56	0.57	10.06	10.03	8.49	7.22
S31	0.69	0.68	9.71	9.51	3.33	3.07
S32	0.22	0.22	10.70	11.21	12.81	12.81
S33	0.10	0.08	11.82	11.51	14.14	14.52
S34	0.50	0.51	8.16	8.16	8.02	9.13
S35	0.39	0.39	11.35	11.31	10.52	10.59
S36	0.24	0.24	10.43	10.14	8.60	8.81
S37	0.70	0.70	7.77	7.57	7.02	7.02
S38	0.58	0.58	10.22	10.56	6.25	6.39
S39	0.51	0.50	7.69	7.62	7.00	7.00
S40	0.31	0.31	8.76	7.47	7.58	7.27
S41	0.18	0.17	10.58	10.40	11.64	11.70
S42	0.00	0.02	13.52	13.30	20.96	20.03
S43	0.80	0.80	7.35	7.38	6.39	6.41

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

S44	0.69	0.68	7.59	7.57	5.53	5.58
S45	0.42	0.42	8.56	8.68	8.31	8.40
S46	0.23	0.29	10.97	10.76	9.15	9.28
S47	0.38	0.42	9.43	9.33	5.65	6.39
S48	0.28	0.26	10.56	10.78	13.96	14.13
S49	0.05	0.05	11.59	11.50	12.07	11.81
S50	0.36	0.39	9.88	9.69	10.63	10.56
S51	0.61	0.61	7.34	7.20	4.15	4.14
S52	0.19	0.16	8.00	8.04	8.46	8.63
S53	0.55	0.57	8.01	8.03	5.71	5.78
S54	0.21	0.19	12.31	12.46	11.45	11.45
S55	0.39	0.43	9.76	10.08	8.39	8.39
S56	0.58	0.58	9.70	9.68	8.53	8.64
S57	0.48	0.48	7.97	8.07	8.97	8.95
S58	0.64	0.65	10.06	10.01	11.57	11.82
S59	0.16	0.16	14.84	14.83	25.77	25.93
S60	0.23	0.22	8.34	8.46	8.70	8.81

		Session 2				
10MWT score		NBWT score	TUG score	FSST score	10MWT score	SEX
RaterA	RaterB	RaterA	RaterA	RaterA	RaterA	
7.00	7.13	0.78	8.60	5.09	7.39	male
7.21	7.10	0.84	7.32	6.00	6.96	female
7.00	7.26	0.75	8.19	5.71	6.44	female
9.89	10.06	0.16	12.59	18.90	9.46	male
5.62	5.52	0.85	7.51	6.39	5.52	male
8.50	8.49	0.41	9.01	9.31	8.55	female
11.84	12.02	0.30	14.09	10.87	12.23	female
7.10	7.13	0.66	7.80	4.62	6.98	female
6.03	5.96	0.70	7.97	6.16	5.82	female
9.96	9.90	0.43	11.15	11.77	9.59	female
13.05	13.25	0.27	12.90	15.85	11.70	male
9.43	9.72	0.43	10.56	7.63	9.52	male
7.45	7.52	0.73	7.22	5.78	7.70	male
7.18	7.62	0.48	7.25	7.77	7.18	male
10.89	10.77	0.30	15.12	14.01	10.73	male
9.78	9.84	0.24	15.11	13.02	8.66	male
6.35	7.28	0.33	9.77	9.11	7.03	male
10.03	9.94	0.28	12.09	13.66	10.01	male
8.62	8.76	0.33	11.07	9.92	8.55	male
6.76	6.90	0.42	9.00	7.32	6.53	female
9.77	9.83	0.46	10.13	9.49	9.62	male
11.88	11.64	0.51	11.87	9.13	11.54	male
7.94	8.01	0.49	10.63	11.93	7.55	male
8.37	8.34	0.42	8.31	9.43	8.81	male
17.26	17.06	0.17	20.11	31.77	18.22	male
10.15	10.27	0.14	15.33	13.83	10.03	male
8.01	7.96	0.33	8.72	8.22	8.13	male
10.45	11.06	0.45	13.80	20.19	10.41	male
7.90	8.15	0.17	10.02	16.32	7.78	male
8.63	8.38	0.52	10.21	7.13	8.22	male
8.20	8.21	0.87	3.56	8.96	8.53	male
9.33	9.33	0.14	11.22	11.76	9.19	male
9.89	9.82	0.27	9.22	14.69	11.52	male
9.19	7.63	0.58	7.09	8.10	8.62	female
9.72	9.69	0.30	9.62	10.28	9.93	male
9.64	9.53	0.28	11.22	9.26	9.25	male
7.96	7.65	0.72	7.88	7.50	6.13	male
10.50	11.03	0.74	8.87	7.31	9.62	male
6.72	6.62	0.55	7.46	8.08	7.08	female
7.27	7.19	0.55	8.87	7.41	6.72	female
8.08	8.03	0.09	10.00	11.09	7.83	female
11.01	11.22	0.02	13.33	17.12	9.96	male
6.41	6.50	0.97	7.46	5.81	6.38	female

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

8.15	8.13	0.79	7.08	6.13	8.15	male
8.60	8.71	0.39	8.08	7.75	8.00	female
8.44	8.40	0.23	9.76	9.81	8.16	male
8.63	8.58	0.46	6.94	5.95	7.56	female
9.13	8.84	0.18	10.66	15.59	9.02	male
8.25	8.06	0.14	11.56	12.39	8.02	female
9.26	9.20	0.43	8.77	10.94	8.44	female
6.83	6.84	0.59	7.25	3.97	7.06	male
7.15	6.96	0.37	7.78	7.70	6.76	male
7.56	7.51	0.84	8.34	5.19	7.65	male
7.32	7.47	0.18	11.13	9.27	7.51	male
7.81	7.84	0.41	8.84	8.34	7.64	male
7.81	7.63	0.37	9.33	8.26	7.88	male
6.60	6.56	0.36	8.59	8.44	7.08	male
8.78	8.75	0.74	9.69	12.27	8.28	male
12.20	12.40	0.21	14.45	21.78	11.96	female
8.08	8.59	0.26	7.35	7.33	8.15	male

AGE	LEVEL	ETIOLOGY	K-LEVEL
29	TT	trauma	3
52	TT	trauma	3
57	TT	trauma	3
71	TF	infection	2
28	TT	trauma	3
39	TF	cancer	3
64	TT	cancer	2
53	TT	cancer	3
29	TT	dysvascular	3
71	TT	trauma	2
67	TT	dysvascular	2
56	TT	dysvascular	3
26	TT	infection	3
51	TT	trauma	3
55	TT	dysvascular	2
60	TF	trauma	2
45	TT	trauma	2
76	TT	trauma	2
35	TF	infection	3
51	TF	trauma	2
59	TT	trauma	3
71	TT	trauma	3
57	TF	trauma	2
58	TT	infection	3
68	TF	infection	2
69	TF	dysvascular	3
62	TT	infection	2
61	TT	dysvascular	2
62	TT	dysvascular	3
48	TT	trauma	3
20	TT	trauma	4
48	TT	dysvascular	3
75	TT	infection	3
51	TT	trauma	3
81	TT	dysvascular	2
72	TF	trauma	3
44	TT	infection	4
26	TT	cancer	3
54	TT	cancer	3
59	TT	infection	3
59	TF	infection	2
84	TT	trauma	2
27	TT	cancer	4

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

55	TT	trauma	4
34	TT	trauma	3
60	TT	infection	3
74	TT	other	2
67	TF	trauma	2
71	TF	trauma	3
45	TF	cancer	3
40	TT	infection	4
18	TF	cancer	4
59	TT	trauma	4
59	TT	infection	3
84	TT	trauma	3
61	TT	dysvascular	3
69	TT	trauma	3
56	TT	trauma	4
64	TF	cancer	3
60	TT	dysvascular	3