Interrater reliability of mechanical tests for functional classification of transtibial prosthesis components distal to the socket

Matthew J. Major, PhD1–2* William Brett Johnson, PhD1 Steven A. Gard, PhD1–2
1Department of Physical Medicine and Rehabilitation, Northwestern University Prosthetics-Orthotics Center, Chicago, IL; 2Jesse Brown Department of Veterans Affairs Medical Center, Chicago, IL

Abstract—Substantial evidence suggests that the design and associated mechanical function of lower-limb prostheses affects user health and mobility, supporting common standards of clinical practice for appropriate matching of prosthesis design and user needs. This matching process is dependent on accurate and reliable methods for the functional classification of prosthetic components. The American Orthotic & Prosthetic Association developed a set of tests for L-code characterization of prosthesis mechanical properties to facilitate functional classification of passive below-knee prosthetic components. The mechanical tests require use of test-specific fixtures to be installed in a materials testing machine by a test administrator. Therefore, the purpose of this study was to assess the interrater reliability of test outcomes between two administrators using the same testing facility. Ten prosthetic components (8 feet and 2 pylons) that spanned the range of commercial designs were subjected to all appropriate tests. Tests with scalar outcomes demonstrated high interrater reliability (intraclass correlation coefficient (2,1) >/= 0.935), and there was no discrepancy in observation-based outcomes between administrators, suggesting that between-administrator variability may not present a significant source of error. These results support the integration of these mechanical tests for prosthesis classification, which will help enhance objectivity and optimization of the prosthesis-patient matching process for maximizing rehabilitation outcomes.

INTRODUCTION

There is a substantial body of literature indicating that the design and associated function of passive lower-limb prostheses affect user performance in terms of walking dynamics, balance, and efficiency [1–3]. Recent investigations have begun to explore and define the fundamental relationships between user performance (e.g., metabolic cost, joint dynamics, and residuum forces) and isolated mechanical properties (i.e., stiffness and damping) of passive prostheses, measured through mechanical characterization tests independent of the user [4–7]. Summarily, the results from these studies strongly suggest that the mechanical function of prostheses has an important role in the health and mobility of the user. Consequently, the common standards of clinical practice that advocate appropriate matching of prosthesis design and user needs and functional potential seems reasonable in order to optimize rehabilitation outcomes. However, the matching process involved in prosthesis prescription guidelines is dependent on accurate and reliable methods

Key words: amputation, below-knee, characterization, function, health, mechanical, mobility, properties, prosthesis, reimbursement.

Abbreviations: AOPA = American Orthotic & Prosthetic Association, ICC = intraclass correlation coefficient.
*Address all correspondence to Matthew J. Major, PhD; Department of Physical Medicine and Rehabilitation, Northwestern University, 680 N. Lake Shore Dr, Suite 1100, Chicago, IL 60611; 312-503-5731; fax: 312-503-5760. Email: matthew-major@northwestern.edu
http://dx.doi.org/10.1682/JRRD.2014.12.0300
for the functional classification of prosthetic components (e.g., multiaxial, dynamic response) that are based exclusively on associated mechanical function rather than appearance or manufacturer claims. Importantly, classifications that are accompanied by information on mechanical function and properties would also lay the groundwork for optimizing prescription guidelines by improving the resolution of methods for clinical judgment and device recommendation.

The American Orthotic & Prosthetic Association (AOPA) began the “Prosthetic Foot Project” in 2007 as a means to develop standardized methods for improving the accuracy and precision of the functional classification of below-knee prosthetic components. This effort resulted in the AOPA Prosthetic Foot Project Report (hereafter referred to as the “Report”) [8]. One of the primary objectives of the Report was that the described testing methods would provide an alternative to the historically used subjective methods of component classification and eventually be adopted as guidelines for the Centers for Medicare and Medicaid Healthcare Common Procedure Code System L-code system, which is used for component reimbursement. The series of designed mechanical tests described in the Report characterize various aspects of prosthesis mechanical properties for passive feet and pylon endoskeletal components, such as range of motion in three planes, vertical displacement, and independent heel and keel stiffness and damping. Functional classification and corresponding L-codes [9] are assigned based on whether test outcomes meet predefined scalar thresholds or observation-based pass/fail criteria.

All of the mechanical tests described in the Report are designed for use with a materials testing machine, and each test requires test-specific fixtures to install and load the prosthetic component as set up by a test administrator. Importantly, these tests are meant to provide standardized procedures that may be implemented by independent facilities with a desire to classify components and access to the required equipment. Because the function, accuracy, and precision of material testing machines are fairly standard, the primary source of error for these tests is the involvement of the test administrator during the multistage setup of test fixtures and prosthetic components. Given that classification is based on strict threshold and pass/fail criteria, minor alterations in relative fixture orientation and positioning of the prosthetic components may generate important differences in outcomes. Consequently, prior to advocating the use and adoption of these characterization tests as a standardized framework for component classification, the reliability of test outcomes between different administrators must be assessed. Therefore, the purpose of this study was to assess the interrater reliability of mechanical test outcomes between two raters using the same testing facility. A secondary objective of this study was to evaluate the reliability of L-code assignment between two test administrators using the guidelines of the Report.

METHODS

The Report describes a series of mechanical characterization tests that apply to both feet and pylons, of which 10 tests generate scalar values to check against thresholds and 3 tests require observation of the components’ behavior during evaluation to confirm whether the component passes or fails to achieve a certain motion. A brief description of these outcome metrics and associated mechanical tests is reported in Table 1, and the full testing procedures are described in the Report [8].

Protocol

Eight prosthetic feet (two each from the following classifications: solid ankle cushion heel, single-axis, dynamic, and multiaxial) and two vertical shock absorbing pylons with torsional adapters were subjected to all appropriate mechanical tests (e.g., pylons were only tested with the axial torque absorption, vertical loading, and dynamic pylon test). These components were donated as new by several prosthetic manufacturers and, if necessary, assembled based on manufacturer specifications. Components were selected in order to test a set of currently prescribed devices that spanned the range of commercial designs and functional classifications [8]. As required by the Report, all tests were performed using components for an “A80” (i.e., 80 kg) patient (and a standard 27 left side component in the case of feet), following the detailed testing procedures, and at normal environmental conditions (temperature, humidity, etc.). A hydraulic-driven materials test machine (model 8800, Instron; Norwood, Massachusetts) was used for all testing, and fixtures were fabricated in-house. Two research engineers independently administered each mechanical test and were responsible for the entire setup of fixtures and component installation, as well as tuning the materials test machine.
Table 1.
Outcome metrics and corresponding mechanical test.

<table>
<thead>
<tr>
<th>Outcome Metric</th>
<th>Pass/Fail Criteria</th>
<th>Mechanical Test and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical Linear Displacement (mm)</td>
<td>Scalar Value Threshold</td>
<td>Keel test—Vertical loading and unloading of foot onto 20° sagittal-plane inclined surface.</td>
</tr>
<tr>
<td>Vertical Linear Displacement (mm)</td>
<td>Scalar Value Threshold</td>
<td>Heel test—Vertical loading and unloading of foot onto 15° sagittal-plane declined surface.</td>
</tr>
<tr>
<td>Coronal-Plane Angular Displacement (°)</td>
<td>Scalar Value Threshold</td>
<td>Multiaxial test—Vertical loading (medial biased plantar force vector) and unloading of foot fixed in cradle with free coronal-plane rotation.</td>
</tr>
<tr>
<td>Transverse-Plane Angular Displacement (°)</td>
<td>Scalar Value Threshold</td>
<td>*Axial torque absorption test—Transverse-plane torque loading and unloading applied to foot and pylon constrained by fixed end and opposing end with free transverse-plane rotation.</td>
</tr>
<tr>
<td>Vertical Linear Displacement (mm)</td>
<td>Scalar Value Threshold</td>
<td>Vertical loading test—Vertical loading and unloading of foot and pylon onto sagittal-plane level surface.</td>
</tr>
<tr>
<td>Vertical Linear Displacement (mm)</td>
<td>Scalar Value Threshold</td>
<td>Dynamic pylon test—Sagittal-plane torque loading and unloading applied to distal end of foot and pylon (i.e., cantilever load scenario).</td>
</tr>
<tr>
<td>Horizontal Linear Displacement (mm)</td>
<td>Scalar Value Threshold</td>
<td>¹Horizontal displacement test (heel)—Vertical loading and unloading of foot angled at 15° dorsiflexion onto sagittal-plane level surface with free sagittal-plane linear translation.</td>
</tr>
<tr>
<td>Heel Contact With Loading Surface</td>
<td>Observation-Based</td>
<td>Single axis test—Vertical loading and unloading of foot onto 15° sagittal-plane declined surface.</td>
</tr>
<tr>
<td>Forefoot Contact With Loading Surface</td>
<td>Observation-Based</td>
<td>Multiaxial test—Vertical loading and unloading of foot onto 8° sagittal-plane declined surface.</td>
</tr>
<tr>
<td>Heel Contact With Loading Surface</td>
<td>Observation-Based</td>
<td>Multiaxial test—Vertical loading and unloading of foot onto 10° sagittal-plane inclined surface.</td>
</tr>
</tbody>
</table>

*For foot testing, foot was constrained in cradle and torque was applied to distal end of foot by suspending known weight across frictionless pulley.

†Mounting fixtures to angle foot were used because International Organization for Standardization test frame was not available as described in Report.

control parameters for each component. Importantly, the test setup was broken down and removed from the materials testing machine by each administrator following the series of tests.

Statistical Analysis

To assess interrater reliability (i.e., level of agreement of outcomes between test administrators), the intraclass correlation coefficient (ICC) was estimated for each test of which a scalar value was produced. The ICC model of two-way random single measures with an absolute agreement condition (ICC(2,1)) was used for this analysis, in which both the test administrator and prosthetic component were considered to be a random sample of the larger population of interest. A Bland-Altman plot, a graphical display of assessor (i.e., test administrator) difference versus mean across components [10], was produced for each test to provide complementary information on level of agreement and error (95% limits of agreement), as well as presence of fixed or proportional bias. Practically, presence of a fixed bias suggests that the measurements from one administrator would consistently be offset compared with those of the other administrator, whereas proportional bias suggests that measurement differences between administrators are dependent on the magnitude of the measurement (e.g., error is expected to be greater with larger and/or smaller values). Fixed and proportional bias were statistically assessed by determining whether the mean error (i.e., difference) between administrators was statistically different than zero (one-sample t-test) and the strength and significance of correlation between administrator average and error through estimation of the Spearman correlation coefficient (ρ), respectively. Given the differences in component designs, many outcome
measure averages across administrators were found to be of nonnormal distribution using the Shapiro-Wilk test with the presence of outliers, and so the nonparametric Spearman $\rho$ was used to estimate correlations between average and error to account for such outliers. The critical alpha for these tests was set at 0.05. For those tests that did not produce a scalar value, a comparison was made of observation-based metrics between both administrators. Following data collection, L-codes were assigned to the components using the independent results of both administrators based on the recommendations as specified in the Report. Since each component was considered a “black box” and not based on manufacturer classification or design specifics, these L-codes are those considered applicable given the test results and not necessarily reflective of final assignment.

RESULTS

Estimates of the interrater ICC values are displayed in Table 2, Bland-Altman plots are displayed in Figures 1–10, and scalar values and applicable L-codes resulting from the mechanical tests for each component as measured by test administrators 1 and 2 are located in the Appendix (available online only). For those tests that required only observation-based measures, no discrepancy was found between test administrators. The Bland-Altman analysis suggested that no significant presence of proportional bias was detected, and the mean error was only significantly different than zero, and hence suggestive of fixed bias, for two outcomes: vertical linear displacement for the heel test (Figure 3; fixed offset = 0.4 mm; $p = 0.03$) and transverse-plane angular displacement for the axial torque absorption test (Figure 6; fixed offset = 0.75°; $p = 0.03$). Only coronal-plane angular displacement of the multiaxial test produced a discrepancy in L-code assignment between Administrators for a single-axis foot, in which code L5986 (all lower-limb prostheses, multiaxial rotation unit [9]) was not applicable based on results for administrator 2 (single-axis foot 1, Appendix).

DISCUSSION

For all tests with scalar outcomes, the ICC values suggest a high level of agreement between administrators [11]. These results indicate that the tasks of setting up test fixtures and installing the prosthetic components by independent administrators do not present a substantial source of error to the measurement outcomes. The fixed bias for the vertical linear displacement and axial torque absorption tests as suggested by the Bland-Altman analyses were minimal and may not be of considerable concern regarding interrater agreement given that ICC values were high for these tests. However, practically, the scalar values from these tests are checked against thresholds to classify components. Consequently, it is possible that a component may be classified differently between test administrators if the fixed bias is sufficient for components to consistently pass a particular threshold. However, this type of fixed bias error can be accounted for by establishing a “zeroing” procedure. For example, a component that yields a known displacement during these tests can be used to obtain a baseline reading and provide a scalar offset for subsequent tests.

Table 2.
Intraclass correlation coefficient (ICC(2,1)) values for scalar metrics.

<table>
<thead>
<tr>
<th>Outcome Metric and Test</th>
<th>ICC Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical linear displacement of keel test ($n = 8$)</td>
<td>0.988 (0.942–0.988)</td>
<td>0.001</td>
</tr>
<tr>
<td>Vertical energy return of keel test ($n = 8$)</td>
<td>0.998 (0.991–1.000)</td>
<td>0.001</td>
</tr>
<tr>
<td>Vertical linear displacement of heel test ($n = 8$)</td>
<td>0.981 (0.815–0.997)</td>
<td>0.001</td>
</tr>
<tr>
<td>Vertical energy return of heel test ($n = 8$)</td>
<td>0.966 (0.842–0.993)</td>
<td>0.001</td>
</tr>
<tr>
<td>Coronal-plane angular displacement of multiaxial test ($n = 8$)</td>
<td>0.988 (0.943–0.998)</td>
<td>0.001</td>
</tr>
<tr>
<td>Transverse-plane angular displacement of axial torque absorption test ($n = 10$)</td>
<td>0.997 (0.977–0.999)</td>
<td>0.001</td>
</tr>
<tr>
<td>Vertical linear displacement of vertical loading test ($n = 10$)</td>
<td>0.935 (0.772–0.983)</td>
<td>0.001</td>
</tr>
<tr>
<td>Vertical linear displacement of dynamic pylon test ($n = 10$)</td>
<td>0.990 (0.951–0.998)</td>
<td>0.001</td>
</tr>
<tr>
<td>Horizontal displacement of heel horizontal displacement test ($n = 8$)</td>
<td>0.997 (0.987–0.999)</td>
<td>0.001</td>
</tr>
<tr>
<td>Horizontal displacement of keel horizontal displacement test ($n = 8$)</td>
<td>0.996 (0.979–0.999)</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Outcome measure results from only one mechanical test, coronal-plane angular displacement of the multiaxial test, produced a discrepancy in L-code assignment between administrators for a single-axis foot. The reason for this discrepancy highlights one of the issues with using scalar value thresholds for L-code assignment, in which the scalar threshold is too similar in magnitude to the actual displacement of the component. In this case, the scalar threshold was 8° and the mean angular displacement of the single-axis foot was 8.2° (7.9° and 8.4° for administrators 1 and 2, respectively). Consequently, as the estimated variability between test administrators suggests that outcome displacement may realistically vary from 6° to 10° (Figure 5), it is conceivable that L-code assignment
may be different between administrators, as demonstrated in this example. Inclusion of measure outcome tolerances may be a useful amendment to scalar thresholds in order to account for this source of error.

This process of improving the functional classification (via L-codes) of prosthetic components through standardized mechanical tests mirrors the recent efforts in improving functional mobility and rehabilitation potential classification.
MAJOR et al. Transtibial prosthesis classification

Figure 9.
Bland-Altman plot for linear displacement of heel horizontal displacement test. Solid black line = difference mean, solid gray line = vertical 0 axis, dashed lines = 95% limits of agreement. SD = standard deviation.

Figure 10.
Bland-Altman plot for linear displacement of keel horizontal displacement test. Solid black line = difference mean, solid gray line = vertical 0 axis, dashed lines = 95% limits of agreement. SD = standard deviation.

Overall, the mechanical tests as described in the Report demonstrated high interrater reliability within the limits of this assessment to suggest that variability between test administrators may not present a significant source of error. Consequently, these results suggest that given appropriate testing equipment, users of these tests can have a high level of confidence that transtibial prosthesis components distal to the socket will not be misclassified because of interrater disagreement. The results from this study emphasize the utility of these mechanical tests to serve an objective, standardized process for characterizing the mechanical properties of prosthesis components for functional classification. Adoption of standardized testing will satisfy an important need of the prosthetics community by minimizing bias and subjectivity and enhancing transparency of the reimbursement of current and future prosthetic devices.

ACKNOWLEDGMENTS

Author Contributions:
Study concept and design: M. J. Major, S. A. Gard.
REFERENCES


Design of test fixtures: M. J. Major.
Acquisition of data: M. J. Major, W. B. Johnson.
Analysis and interpretation of data: M. J. Major, S. A. Gard.

Financial Disclosures: The authors have declared that no competing interests exist.

Funding/Support: This material was based on work supported by the AOPA (RFP 083112).

Additional Contributions: We would like to thank Dilip Thaker and Edward Grahn for their assistance in the design and fabrication of the test fixtures and Jonathon Naft for his insightful discussions on this topic and constructive feedback during preparation of the manuscript.


Submitted for publication December 1, 2014. Accepted in revised form February 25, 2015.

This article and any supplementary material should be cited as follows:

ResearcherID: Matthew J. Major, PhD: E-7372-2012; William Brett Johnson, PhD: K-5231-2012; Steven A. Gard, PhD: D-9935-2011