Reliability of a Single-Session Isokinetic and Isometric Strength Measurement Protocol in Older Men

T. Brock Symons,1 Anthony A. Vandervoort,1,2 Charles L. Rice,1 Tom J. Overend,2 and Greg D. Marsh1

Canadian Centre for Activity and Aging,1 Schools of Kinesiology and Physical Therapy, Faculty of Health Sciences, The University of Western Ontario, London, Canada.

Background. The purposes of the current study were (a) to determine the test–retest reliability of a single-session isokinetic and isometric strength testing protocol in older healthy men, and (b) to compare the outcomes of the reliability measures derived from averaged torque scores with those derived from a single peak torque score.

Methods. In 19 men (mean age, 72 ± 5 years), both lower limbs were assessed independently on 2 separate test days using the Biodex System 3 dynamometer. After completing a 5-minute warm-up, each man performed three submaximal knee extensions followed by five maximal contractions at 90°/s (CON), 0°/s (ISO), and −90°/s (ECC). Average (best 3 of 5) and peak CON, ISO, and ECC torque, and CON work and CON power were determined. Peak CON work and CON power were recorded from the highest peak torque concentric contraction (HPTCC).

Results. Intraclass correlation coefficients ranging from 0.84 to 0.94 were found to have good reliability. The typical error as a coefficient of variation ranged from 8% to 10% for averaged measures and from 8% to 17% for peak torque and HPTCC. The ratio limits of agreement for average and peak CON, ISO, and ECC torque ranged from 23% to 33% and from 40% to 54% for average CON and HPTCC work and power.

Conclusions. The test–retest reliability of a single-session isokinetic and isometric strength testing protocol in this group of older healthy men displayed good relative reliability (intraclass correlation coefficient > 0.84); however, because the typical error as a coefficient of variation and ratio limits of agreement (absolute reliability) were large, single-session testing is not recommended.

Many protocols involving concentric (muscle shortening) and to a lesser extent eccentric (muscle lengthening) muscle actions have been found to be reliable (consistency or reproducibility of measurements [1–3]) in healthy young persons when isokinetic knee extensor torque is measured (4–8). However, only a few studies have assessed the reliability of isokinetic measurements in older adults. Capranica and colleagues (9) evaluated the test–retest reliability of isokinetic strength testing in the knee extensors and flexors of older women (mean age, 68 years) for concentric torque, power, and work. Frontera and colleagues (10) assessed the test–retest reliability of isokinetic concentric elbow and knee extension and flexion strength in both older women and men (mean age, 60 years) at two different angular velocities. In addition to the limited number of reliability studies involving older men, isometric torque, eccentric torque, and the Biodex Multi-Joint System 3 dynamometer (Biodex Medical, Shirley, NY) reliabilities have not been studied in this population. In addition, with the recent findings that eccentric resistance training resulted in significant improvements in strength, balance, and stair descent (11), eccentric strength testing and training will become more prevalent.

The capacity to minimize the amount of measurement error, the positive or negative deviations between the true score and the observed score, allows researchers and clinicians to have confidence in their data collection and permit lucid conclusions to be drawn from the data (3). Reliable measurement protocols ensure that observed changes in performance over time reflect real gains and are not merely artifacts of the measurement or procedure (10).

The reliance on multiple test sessions to enable participants to become fully comfortable and familiar with the testing protocol and apparatus is not always possible or practical. Multiple test sessions are not always feasible in research or clinical settings involving large numbers of participants for many reasons, including time requirements, cost, and availability of equipment and facilities (12). Therefore, the purposes of the current study were (a) to determine the test–retest reliability of a single-session isokinetic and isometric strength testing protocol in older healthy men, and (b) to compare the outcomes of the reliability measures derived from averaged torque scores with those derived from a single peak torque score.

Methods

Participants

Nineteen healthy men (age, 72 ± 5 years; height, 173 ± 7 cm; weight, 83 ± 11 kg [mean ± standard deviation]) participated in the study. All were free of any cardiovascular and lower limb musculoskeletal or neuromuscular limitations. All participants gave written informed consent, and the university’s Research Ethics Committee approved all procedures.

Equipment and Participant Positioning

All strength testing was performed on the Biodex Multi-Joint dynamometer and assessed using Biodex System 3.
Advantage Software, version 3.2. Before the start of each testing session, the Biodex was calibrated in accordance with the manufacturer’s specifications. For all strength testing protocols, the cushion (the deceleration point) was set at “high” and the attachment sensitivity (the acceleration control for the knee attachment) was factory set. Participants sat in a comfortable upright position with the Biodex seat back tilted to an angle of approximately 85°, resulting in a hip angle of approximately 110° flexion. The participant was stabilized with 2 shoulder straps, a waist strap, and a thigh strap. The rotational axis of the knee (lateral femoral epicondyle) was aligned with the center of the dynamometer shaft. Adjustments were made to the length of the knee attachment to ensure that the ankle strap was proximal to the lateral and medial malleoli and comfortable for the participant. The participant’s range of motion was established about the knee joint (0° of full extension to −93° flexion) (13). Any further chair or participant positioning was performed at this time and all chair settings were recorded to ensure setup reproducibility on subsequent visits. Lower limb weight was determined to negate the influence of gravity on all torques measured in accordance with the manufacturer’s specifications.

Test Protocol

The same investigator administered all strength tests and provided all verbal instructions. Strength testing was performed on both the dominant and nondominant lower limbs, and the starting limb was randomized. We maintained the order of isokinetic strength testing for all participants as follows: concentric (CON) at 90°/s were performed first, followed by 5-second isometric (ISO) at 0°/s (=90° knee angle), and finally eccentric (ECC) at −90°/s contractions. Each man participated in two identical test sessions (test day 1 and test day 2) separated by 2–10 days, and the tests were conducted at approximately the same time each day. The number of days between test days was not standardized because it is not always feasible to control for such a variable when testing large numbers of persons or in a clinical setting. Before the start of strength testing, all participants performed a 5-minute warm-up on a stationary cycle ergometer at a low 1 kilopond load.

Once positioned in the Biodex chair, the participant was given a verbal description of the contraction to be performed. Furthermore, his lower limb was moved passively through the desired range of motion by the investigator, who also explained the resistance pattern of the dynamometer arm. Participants then performed three submaximal contractions at approximately 50%–65% of their perceived maximal effort. After an additional brief description of the contraction to be performed (i.e., to exert a maximal effort by contracting as hard and as fast as possible), the participant performed five maximal voluntary contractions. Rest periods of 5 seconds were given between each contraction, during which time the leg passively returned to the starting position. Five seconds provided a sufficient recovery between contractions because no decrease in peak torque was observed across the five contractions. A 15-second rest period was given between the ISO contractions because they are considered to be more taxing; however, because we noted no sign of fatigue during the ISO test, a longer period was not justified. We added a 2-minute rest period before the start of the next strength test. Throughout the session, each participant was given consistent verbal encouragement and visual feedback via the computer monitor of the Biodex.

Data and Statistical Analyses

We recorded the average peak torque (the best three of five contractions) and the highest peak torque (best contraction of the five performed) for CON, ISO, and ECC strength. We determined average CON work and CON power as above, whereas peak CON work and CON power were recorded from the highest peak torque concentric contraction (HPTCC).

For the purpose of statistical analyses, we combined data from the dominant and nondominant lower limbs to provide a sample size of 38. We did this because sample sizes of fewer than 30 participants are not considered an adequate representation of a normal distribution (3). When we combined our data, the sample tested did become more homogenous. However, typical error can be predicted from a homogeneous sample or a sample with only a few participants from a specific population, and we can assume that the estimated typical error applies to any member of the specific population being tested (2). In addition, ICCs are sensitive to the heterogeneity of the sample (2). A homogenous sample with similar scores will most likely result in a lower ICC estimate than a more heterogeneous sample with a wide range of scores, which will most likely generate a greater ICC (14). Also, if a statistical difference had been found between the dominant and nondominant lower limbs, this would probably have required a separate analysis for each lower limb and the development of separate statistical measures for each dominant and nondominant lower limb because they were significantly different.

Furthermore, we generated Bland-Altman plots (15) of the difference between test day 1 (TD1) and test day 2 (TD2) versus the mean of TD1 and TD2 for each participant using the raw data scores for all measures (Figures 1 and 2). We determined heteroscedasticity (nonuniform scatter) by visual inspection of the Bland-Altman plots and we deemed them to be present when the difference scores for participants at one end of the plot demonstrated a tendency for larger values (2). Furthermore, Nevill and Atkinson (16) concluded that heteroscedastic errors are the norm when agreement between variables recorded on a ratio scale (strength or power) is evaluated, and they support the use of log-transformed data, thus stabilizing the variance and normalizing the distribution. To maintain precision, we multiplied the natural logarithm by 100.

We used the following statistical measures to determine test–retest reliability for all measures. We analyzed reproducibility bias using paired t test analysis between TD1 and TD2 (p < .05), and we determined the percentage change in the mean between TD1 and TD2. With intraclass correlations (ICC2,1 and ICC2,3), the first subscript number denotes the “model” and the second subscript number signifies the “form.” Model 2 was chosen because it is based on a repeated measures analysis of variance and the same rater assessed all participants. Form 1 represents the use of a single score, whereas form 3 represents the use of...
a mean of three scores (3). We used ICC2,1 to analyze the highest peak torque contraction scores and the HPTCC work and power scores, whereas we used ICC2,3 to analyze the CON, ISO, and ECC average peak torque scores and average CON work and power scores. We calculated typical error (variation in a participant’s score from measurement to measurement) expressed as a coefficient of variation (CVTE) in accordance with the method of Hopkins (2). We determined ratio limits of agreement (RLOA), which reflect the 95% probability limits between which the difference score between any two tests should lie, in accordance with the method of Atkinson and Nevill (1).

\[
\text{RLOA} = \text{antilog}\left(\frac{\text{SD}(\text{TD1-TD2})}{1.96}\right)
\]

and

\[
\text{CV}_{\text{TE}} = \left(\text{antilog}\left(\frac{\text{SD}(\text{TD1-TD2})}{\sqrt{2}}\right)\times 100\right) - 100
\]

where

- \(SD = \text{standard deviation}\)
- 1.96 = the 95th percentile
- TD1-TD2 = the difference between test day 1 and test day 2

Statistical analyses were performed using SPSS 9.0 (SPSS, Chicago, IL) and Microsoft Excel (Microsoft, Seattle, WA).

RESULTS

Table 1 presents the means and standard deviations for all isokinetic measures. As expected, CON, ISO, and ECC torques varied in accordance with the force velocity curve, with CON torque giving the lowest values, ISO torque falling in the middle, and ECC torque yielding the highest values. Isometric torque, both averaged and peak values, demonstrated a small but statistically significant mean bias, with TD2 scores increasing by 7.4%. This was also the case for average CON work (7.7%) and HPTCC work (7.1%). Average CON power also showed a mean bias with increased power scores on TD2; however, HPTCC power showed no bias. These fluctuations in mean values had little effect on the intraclass correlations for averaged and peak CON, ISO, and ECC torque, which are presented in

![Figure 1. Bland-Altman plots of averaged and peak torque concentric (CON), isometric (ISO), and eccentric (ECC) muscle actions. The long dashes represent upper and lower 95% limits of agreement. The short dashes represent the mean bias.](image-url)
Tables 2 and 3. They ranged from 0.87 to 0.93 for all average variables and from 0.84 to 0.91 for all peak and HPTCC values.

Tables 2 and 3 show CV TE for all measures. Because the variations between TD1 and TD2 are larger than 5%, we expressed CVTE and RLOA values as factors to be more accurate (2). The percentage of random variation in a participant’s score from TD1 to TD2 for averaged and peak CON, ISO, and ECC torque ranged from $3/4^1.08$ to $3/4^1.11$. Averaged CON and peak CON torque show the greatest within-participant variation, whereas averaged ECC and peak ECC torque resulted in the lowest CV TE values. The RLOA for averaged CON, ISO, and ECC torque and CON work and power ranged from $3/4^1.23$ for averaged ECC torque to $3/4^1.42$ for averaged CON power.

**DISCUSSION**

Relative reliability (the degree to which individuals preserve their rank in a sample through repeated measures [17]) scores for CON, ISO, and ECC torque were good, with ICCs greater than 0.89 (3). Concentric work and CON power displayed slightly lower ICCs, ranging from 0.84 to 0.89. Within-participant changes from day-to-day measurements or absolute reliability (the degree to which a person’s observed score varies with repeated measures [17]) showed greater ranges across these variables. The use of average torque scores versus single peak torque scores failed to produce different results for any measures.

Paired t tests performed on the natural log-transformed data revealed a small bias toward higher scores during the second test day for 5 of the 10 isokinetic measures. The mean change between two trials can be attributed to random error, systematic bias (1,2), or both. Random error is due to chance (sampling error) and is unpredictable, either increasing or decreasing the observed score (2,3). Systematic bias is predictable (nonrandom), either overestimating or underestimating the mean in one direction (1–3). Random error can occur because of unpredictable factors such as

Table 2. Descriptive Statistics for Isokinetic and Isometric Measures for Older Men (Mean ± SD + % Mean Difference)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Average Values</th>
<th>Peak Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test Day 1</td>
<td>Test Day 2</td>
</tr>
<tr>
<td>CON (Nm ± SD)</td>
<td>113.0 ± 31.1</td>
<td>116.3 ± 27.8</td>
</tr>
<tr>
<td>ISO (Nm ± SD)</td>
<td>147.8 ± 36.9</td>
<td>157.5 ± 34.5*</td>
</tr>
<tr>
<td>ECC (Nm ± SD)</td>
<td>177.7 ± 39.3</td>
<td>183.2 ± 38.6</td>
</tr>
<tr>
<td>CON Work (J ± SD)</td>
<td>117.5 ± 33.6</td>
<td>124.2 ± 30.5*</td>
</tr>
<tr>
<td>CON Power (W ± SD)</td>
<td>90.0 ± 30.0</td>
<td>95.9 ± 27.9*</td>
</tr>
</tbody>
</table>

Notes: * Significant difference between Test Day 1 and Day 2 ($p < .05$); statistical analysis based on natural log-transformed data.

CON = Concentric; ISO = isometric; ECC = eccentric; % Diff = percent mean difference; SD = Standard deviation.
mechanical variations in the measurement device and mistakes by the investigator or participant (1,3). The most prevalent example of systematic bias is the learning effect or habituation, with participants performing better on a second trial simply because they have benefited from the experience of the first trial (1,2). Other examples of systematic bias include the failure to allow for sufficient recovery between tests, resulting in fatigue, and training, or intervention effects (1–3). In the current study, we noted systematic bias; however, this was not unexpected with the population being examined. It is not surprising that some degree of habituation occurred, because none of the participants had any previous experience performing contractions on a computer-controlled dynamometer and, indeed, most of them had no experience at all with any resistance training exercises.

Relative reliability was high for CON, ISO, and ECC average and peak torque values, having ICCs of 0.89–0.93. According to the suggested guidelines of Portney and Watkins (3), intraclass correlations greater than 0.75 are considered indicative of good reliability. Furthermore, ICCs for average and HPTCC work and power also showed good reliability (0.84–0.89). In the only comparable study of reliability in older adult men, Frontera and colleagues (10) presented their data using Pearson correlation coefficients for different testing protocols, computer-controlled dynamometers, and populations (2). The current study yielded 8% (×/√1.08) to 17% (×/√1.17) variation between trials for all measures. Unfortunately, we found no other studies using CTE for knee extensors in older men after an intensive Internet search. This makes it difficult to determine whether these values represent an acceptable amount of error; however, Hopkins (18) suggested that CTE for most performance tests should be within 1%–5%. When measuring gross changes in knee extensor strength of older men, a somewhat higher CTE would not be too problematic because changes from before to after training are normally large. Nevertheless, the lowest possible CTE is ideal, and it is clear that the trial-to-trial ranges in the current study were greater than desirable. The pretest verbal instructions and the pretest warm-up should be modified. Perhaps more submaximal contractions and possibly one or two maximal contraction warm-up procedures should be added to the pretest warm-up. However, the addition of one or two familiarization sessions before the first day of testing would best reduce the trial-to-trial discrepancy.

The RLOA comprise the second measure of absolute reliability and represent the range within which a participant’s difference score (TD1–TD2) would occur most of the time (2). Therefore, assuming that the errors are normally distributed, 95% of the participant’s difference scores should lie within plus or minus two standard deviations of the mean of the difference scores (3). Like CTE, the limits of agreement should be presented as a ratio because of the heteroscedastic data (1). Data from the current study revealed RLOA values ranging from 23% (0.93 ×/√1.23) to 54% (0.94 ×/√1.54). Consequently, any difference between the two tests of the previously mentioned isokinetic measures should differ by no more than the corresponding percentages above and below the mean bias. The percentages above and below the mean bias for all measures observed were again larger than desired. Thus, any potential increases in strength, work, and power in subsequent visits would have to be large to detect any change caused by training or treatment, and such precision in a testing protocol is not acceptable. Therefore, both measures of absolute reliability (CTE and RLOA) indicate that the single-session testing protocol was not sufficiently reliable to use in research, rehabilitation, or clinical settings in older men, and familiarization sessions are needed to achieve a more precise testing protocol.

Reliability measures derived from averaging the three best contractions of the five performed did not differ from those obtained from the highest peak torque or from the HPTCC (Table 2 and 3). The greatest difference occurred between average CON power and HPTCC power with a 0.05%, 4%, and 12% differences for ICC, CTE, and RLOA, respectively.

Summary

Relative reliability (ICCs > 0.84) for the single-session isokinetic and isometric strength measurement protocol was

---

**Table 2. Reliability Measures for Concentric, Isometric, and Eccentric Torques**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Average Torque</th>
<th>Peak Torque</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CON</td>
<td>ISO</td>
</tr>
<tr>
<td>Intraclass correlation coefficient (ICC)</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Typical error as a coefficient of variation (CVTE) (%)</td>
<td>10.92</td>
<td>7.71</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>8.90</td>
<td>6.29</td>
</tr>
<tr>
<td>Ratio limits of agreement (% of mean)</td>
<td>33.27</td>
<td>22.85</td>
</tr>
</tbody>
</table>

**Note:** CON = concentric; ISO = isometric; ECC = eccentric.

**Table 3. Reliability Measures for Concentric Work and Power**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Average</th>
<th>HPTCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>Power</td>
<td>Work</td>
</tr>
<tr>
<td>Intraclass correlation coefficient (ICC)</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Typical error as a coefficient of variation (CVTE) (%)</td>
<td>12.88</td>
<td>13.43</td>
</tr>
<tr>
<td>Lower confidence limit</td>
<td>10.50</td>
<td>10.95</td>
</tr>
<tr>
<td>Upper confidence limit</td>
<td>16.66</td>
<td>17.38</td>
</tr>
<tr>
<td>Ratio limits of agreement (% of mean)</td>
<td>39.90</td>
<td>41.80</td>
</tr>
</tbody>
</table>

**Note:** HPTCC = Highest peak torque concentric contraction.
highly reliable when older men were examined. However, the CV_{TE} and RLOA (absolute reliability) were larger than expected, and in most research or clinical settings, these values should be smaller. Therefore, we do not recommend the use of a single-session test protocol; one or two familiarization sessions should be added.

ACKNOWLEDGMENTS

Supported by a Pharmacia/Pfizer Research Grant for Physical Activity and Musculoskeletal Injury Prevention from the American College of Sports Medicine Foundation and the Natural Sciences and Engineering Research Council of Canada.

Address correspondence to Anthony A. Vandervoort, PhD, Schools of Kinesiology and Physical Therapy, University of Western Ontario, Room 1400, Elborn College, 1201 Western Road, London, Ontario, Canada N6G 1H1. E-mail: vandervo@uwo.ca

REFERENCES


Received July 9, 2003
Accepted August 22, 2003
Decision Editor: John E. Morley, MB, BCh