Intrarater and interrater reliability of the Anteromedial Reach Test in healthy participants

Background: The Anteromedial Reach Test is a performance-based outcome measure for evaluating dynamic knee stability in patients with anterior cruciate ligament injury. No previously published study has adequately evaluated intrarater or interrater reliability of the Anteromedial Reach Test, so the purpose of this study was to assess these measurement properties in healthy participants prior to their investigation in patients with anterior cruciate ligament injury.

Methods: Two raters (A and B) tested 39 healthy university staff and students (20 men, 19 women). For the intrarater reliability investigation, rater A tested participants on three separate test occasions (days 1, 2, and 3) at the same time of day. For the interrater reliability investigation, raters A and B independently tested participants on the same test occasion (day 3).

Results: There was no significant systematic bias between test occasions or raters. Values of the intraclass correlation coefficient (2,1) were 0.96 for intrarater reliability of both the dominant leg and nondominant leg and 0.97 (dominant leg) and 0.98 (nondominant leg) for interrater reliability. Values for the standard error of measurement were 1.46 (dominant leg) and 1.62 (nondominant leg) for the intrarater investigation, and 1.26 (dominant leg) and 1.04 (nondominant leg) for the interrater investigation. At the 90% confidence level, the minimum detectable change was 3.8% and the error in an individual’s score at a given point in time was ±2.7%.

Conclusion: The Anteromedial Reach Test demonstrated excellent intrarater and interrater reliability in healthy participants. This provides a basis for future investigation of the measurement properties of the Anteromedial Reach Test in patients with anterior cruciate ligament injury.

Keywords: anterior cruciate ligament, injury, dynamic stability, rehabilitation, outcome measures
the English National Health Service,\textsuperscript{14} at an average cost of £1,500 per patient.\textsuperscript{15}

Whether treated surgically or conservatively, several months of rehabilitation are usually required following ACL injury.\textsuperscript{16} Rehabilitation focuses on improving neuromuscular function to enable knee stabilization during dynamic activities, particularly those involving DLEV,\textsuperscript{17} with progress monitored using patient-reported questionnaires, eg, the Lysholm score,\textsuperscript{18} or performance-based measures of physical function, eg, isokinetic strength\textsuperscript{19} and hop testing.\textsuperscript{20} Given that the goal of rehabilitation is often to return patients to athletic activity, it is proposed that performance-based measures are particularly important.\textsuperscript{21} However, many performance-based measures have been criticized for testing the ability to produce force in the sagittal plane, rather than the ability to stabilize the knee during functional, multiplanar motions such as DLEV.\textsuperscript{22,23} To address this criticism, a new ACL-specific, performance-based measure, the Anteromedial Reach Test (ART), was created.\textsuperscript{24,25} Simple and inexpensive, the ART requires patients to stand on one leg while reaching to the anteromedial direction. It aims to test an individual’s ability to dynamically stabilize the knee during DLEV.

Although similar to the anteromedial component of the Star Excursion Balance Test (SEBTam),\textsuperscript{26} the ART is designed to maximize knee involvement. During the SEBTam, participants are allowed to lean backwards and contact the floor with the toes of the reaching foot.\textsuperscript{27} By utilizing this tactic, a knee-injured patient could achieve a maximal reach distance whilst minimizing motion at the knee (Figure 2). In contrast, the ART does not permit participants to lean backwards and requires that they contact the ART board with the heel of the reaching foot.\textsuperscript{25} Therefore, greater knee motion (and possibly muscular activity)\textsuperscript{28} should be required to achieve a maximal distance on the ART (Figure 3). This might explain why, in a preliminary investigation of the measurement properties of the ART, significant differences were detected between the injured and uninjured legs of 30 ACL-deficient patients,\textsuperscript{24} providing evidence of known-groups validity.\textsuperscript{29} Conversely, in a study by Herrington et al, the SEBTam did not detect significant differences between the injured and uninjured legs of 25 ACL-deficient patients, nor did it detect significant differences between these patients and a group of matched healthy controls.\textsuperscript{9}

To be useful in clinical practice, the ART must demonstrate adequate measurement properties, including reliability, validity and responsiveness.\textsuperscript{29} Reliability is often investigated first, being a prerequisite for the other properties.\textsuperscript{30} In addition, it is recommended that initial studies of a new measure are conducted with healthy volunteers rather than patients, to exclude any variability
due to fluctuation in symptoms. For performance-based measures, where measurements are taken by a rater, both intrarater and interrater reliability are important. Intrarater reliability is the extent to which measurements taken by the same rater are consistent, while interrater reliability is the extent to which measurements taken by different raters are similar.

In the aforementioned preliminary investigation of the measurement properties of the ART, intrarater reliability was found to be excellent in healthy volunteers (intraclass correlation coefficient [ICC] 0.96); however, the authors neglected to normalize reach distances for leg length, intrarater reliability was not evaluated, and the report lacked sufficient detail because it was only published in the form of a short conference abstract. Therefore, no previously published study has adequately evaluated intrarater or interrater reliability of the ART, and so these properties require investigation. Also, the possibility of sex-related and bilateral differences in reliability should be considered. For example, significant fluctuations in neuromuscular function can occur throughout the female menstrual cycle, which could increase variability between repeated tests. Additionally, differences in reliability between the dominant leg and nondominant leg have been demonstrated for some performance-based measures (eg, jump testing). Accordingly, the primary objective of this study was to evaluate intrarater and interrater reliability of the ART in healthy participants. Secondary objectives were to evaluate reliability for each sex (men and women) and leg (dominant and nondominant).

Materials and methods
Study design
A repeated-measures design was used. Intrarater reliability was evaluated by comparing ART scores taken by the same rater (rater A) on three separate test occasions (labeled days 1, 2, and 3), a minimum of 2 days and a maximum of 7 days apart. Three test occasions were employed to allow for the possibility of a learning effect between days 1 and 2. In this event, day 1 would be considered a familiarization day, with reliability calculated using data from days 2 and 3 only. Interrater reliability was evaluated by comparing ART scores taken by two different raters (raters A and B) on the same test occasion (day 3). This took place on the final test occasion, so that interrater reliability could be analyzed independent of any learning effect.

Participants
A power calculation determined that 19 participants of each sex were required for a reliability analysis involving two time points or raters, to distinguish $\rho_{0.7}$ from $\rho_{0.9}$ at $\alpha=0.05$ and $\beta=0.2$. Allowing for a 10% dropout rate, 42 volunteer healthy staff and students were recruited from one department in a university in the UK. Participants provided written informed consent and the study was approved by the Nursing and Physiotherapy Ethics Panel (School of Health and Population Sciences, University of Birmingham). Two women and one man subsequently withdrew from the study before its completion, due to scheduling difficulties (n=1), leaving 39 participants (20 men, 19 women). Participant characteristics are shown in Table 1.

Participants were excluded if reporting a history of injury to the legs or lumbar spine, or any balance, neurologic, or uncorrected vision disorders, since these could affect neuromuscular performance. Those aged over 45 years were also excluded because knee proprioception declines with increasing age and incidence of osteoarthritis.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mean (± SD) participant characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men (n=20)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>24.7±4.6</td>
</tr>
<tr>
<td>Height (m)</td>
<td>1.8±0.1</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>77.5±9.6</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>23.9±2.5</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; BMI, body mass index.
To facilitate generalizability of findings and comparison with other studies, participant activity levels were recorded using the Marx Activity Rating Scale, which records the frequency of participation in athletic activity involving running, pivoting, cutting, and deceleration. A mean (± standard deviation) score of 9.7±3.9 was obtained, which is approximately equivalent to playing a sport involving all of these activities once a week (scores eight points), in addition to jogging three or more times a week (scores an additional two points).

Raters

The ART is intended for clinical use by practitioners with varying levels of expertise at using the measure. We therefore selected a rater who had used the ART previously (rater A, a physiotherapy lecturer with 6 years of clinical experience and 2 months of ART experience) and a rater who had no previous experience with the ART (rater B, a physiotherapy lecturer with 7 years of clinical experience). Neither rater had previous experience of using the Star Excursion Balance Test (SEBT). Rater B was familiarized with the ART during a single 30-minute session, prior to commencing the study. Rater A explained and demonstrated the ART procedure to rater B, using a standardized set of instructions. Rater B then practiced administering one bout of the ART, with rater A acting as the participant.

Procedure

The procedure is shown in Figure 4. Participants attended three test occasions (days 1, 2, and 3), barefooted and wearing shorts and a t-shirt. The mean (± standard deviation) interval between days 1 and 2 was 3.9±1.9 days, with 4.8±2.0 days between days 2 and 3. All three test occasions took place at the same time of day. To avoid impairment of neuromuscular function, participants were requested to attempt their normal amount of sleep on nights prior to testing, and avoid vigorous exercise for 24 hours before testing, alcohol or caffeine consumption on the days of testing, and the consumption of food or beverages, other than water, for 2 hours before testing.

Day 1

Testing was administered by rater A. Leg length was measured in the supine position with a standard tape measure, from the anterior superior iliac spine to the distal point of the medial malleolus. The dominant leg was determined to be the leg with which participants would choose to kick a ball. For this, and all subsequent testing, the ART procedure was first explained and demonstrated by the rater, using a standardized set of instructions. Next, participants performed one bout of the ART as previously recommended (eight practice trials, followed by five recorded trials, on each leg). There was 15 seconds between trials for data collection and 5 minutes between legs to avoid fatigue. Leg testing was reassigned according to a counterbalanced, randomized ordering across consecutive participants. Participants maintained their assigned order throughout the study.

Day 2

Testing was administered by rater A, who was blind to previous findings. Participants performed one bout of the ART (eight practice trials, followed by five recorded trials, on each leg).

Day 3

Testing was administered by raters A and B.

Figure 4 Schematic diagram of study design.

Abbreviation: ART, Anteromedial Reach Test.
Day 3
Participants performed two bouts of the ART, with one bout administered by rater A (who was blind to previous findings) and one by rater B. Rater testing was preassigned according to a counterbalanced, randomized ordering across consecutive participants. Raters were blind to each other’s findings. The first bout of the ART comprised eight practice trials, followed by five recorded trials, on each leg. After 5 minutes of rest, participants performed the second bout of the ART (one practice trial, followed by five recorded trials, on each leg). The additional practice trial was to prevent a performance decrease resulting from the rest period.26

ART procedure
The ART procedure has been described previously.25 Participants stood on the plastic ART board (Figure 5), measuring 150 cm × 90 cm. This was marked with four lines (left oblique, right oblique, transverse, sagittal), intersecting at a common origin. Strips of 2.5 cm wide semitransparent masking tape were placed over both oblique lines, which were oriented at 45° angles to the transverse line. New tape was applied for each bout of the ART.

Stance foot positions are shown in Figure 5. Participants were asked to reach as far as possible with the contralateral leg, along the corresponding oblique line, make a single, light touch-down onto the tape with the reaching heel (Figure 6), and return to the double-legged starting position, without moving the stance foot or placing substantive weight through the reach leg, as judged by the rater. This latter requirement was judged to have not been met if the reaching heel contacted the ART board in a heavy, uncontrolled manner, or if body weight was transferred forward onto the reaching heel after making contact with the board. Participants were also required to keep their hands on their hips, not lean backwards, and hold the knee and ankle of the reach leg in maximum extension and dorsiflexion, respectively. If these criteria were not met for any recorded trial, the trial was discounted and repeated.

Following a successful touch-down, the participant maintained contact until a ruler was slid to the back of the reaching heel. The masking tape was then marked at this point with a pencil, and a standard tape measure was used to measure the distance from the origin, while the participant looked away. The mark was then erased.

ART score calculation
Normalized ART scores were calculated as the mean reach distance of five recorded trials, divided by leg length and multiplied by 100.25

Statistical analysis
Statistical analysis was conducted using PASW Statistics (version 18.0.3; IBM, Somers, NY, USA) with the level of statistical significance set a priori at 0.05. Separate analyses were conducted for each leg (dominant leg and nondominant leg) and sex grouping (men, women, both sexes). Data were analyzed in four stages as follows.

Outliers and normality of data
Since outliers can markedly affect reliability statistics,40 they were excluded using a previously reported method for
reliability studies. However, this method for identifying outliers assumes that data are normally distributed. Q-Q plots were used to check normality of the data in preference to statistical tests of normality because the latter are sensitive to outliers even when the underlying distribution is normal. There was no evidence that the distribution of the data departed from normality.

The method used for outlier identification and exclusion was as follows. A participant’s data for two consecutive bouts of the ART were excluded from analyses when the difference between these two bouts lay outside of a calculated 99% acceptance range (mean difference between bouts for the group ± 2.576 standard deviations of the difference). This is because such a large difference between bouts, lying outside the range in which 99% of differences are expected to lie, could be expected to result from error in performing or administering one of the bouts. For the intrarater analysis, acceptance ranges were calculated for differences between days 1 and 2, and days 2 and 3. For the interrater analysis, these were calculated for differences between raters A and B on day 3.

Whether considering both sexes together, or each sex separately, the same outliers were identified. Two male participants exceeded acceptance ranges for intrarater analysis of the dominant leg, so were excluded from this analysis. One of these participants also exceeded the ranges for interrater analyses of both legs, so was excluded from the interrater analyses.

Following the exclusion of outliers, data for use in subsequent analyses were tested for normality using the Shapiro-Wilk test. Because there was no evidence of departure from normality, parametric statistical tests were used in all subsequent analyses.

**Systematic bias**

Systematic bias is a trend for all participants’ scores to improve or worsen between repeated assessments (eg, due to a between-session learning effect). For the intrarater analysis, we used repeated-measures analyses of variance to test for systematic bias between days 1, 2, and 3. Based on the results, data from all three test occasions were used in subsequent intrarater analyses, effectively increasing the sample size and improving the precision of results. For the interrater analysis, we used paired t-tests to test for systematic bias between raters A and B.

**Reliability**

For both intrarater and interrater reliability, we calculated an ICC (2,1) with absolute agreement and a 95% confidence interval. The ICC is the most commonly used reliability index for continuous data. An ICC ≥ 0.7 indicates “good” reliability and is reported as sufficient for using a measure in research. An ICC ≥ 0.9 indicates “excellent” reliability and is reported as sufficient for making clinical decisions regarding individuals.

**Measurement error**

For both the intrarater and interrater analyses, we estimated the standard error of measurement (SEM) as the square root of the mean square error term from the analysis of variance produced during the ICC calculation. A 95% confidence interval for the SEM was calculated using the method of Stratford and Goldsmith. The SEM represents the amount of error associated with a measure, expressed in actual units of measurement.

Using the SEM from the intrarater analysis, the minimum detectable change at the 90% confidence level (MDC90) was estimated as: SEM × 1.64. This is the smallest change in an individual’s score considered to be a true change and not measurement error. Additionally, we estimated the error in an individual’s score at a given point in time at the 90% confidence level as: ± SEM × 1.64. We used the 90% (rather than 95%) confidence level when estimating these values, based on the rationale that an individual’s score should be interpreted more liberally than group scores. An explanation of the clinical application of these values is provided in the Discussion section.

**Results**

**Intrarater analysis**

Mean (± standard deviation) ART scores for days 1, 2, and 3 are shown in Table 2. Repeated-measures analyses of variance indicated no significant systematic bias between testing occasions (Table 2). Reliability and measurement error statistics are presented in Table 3. ICC values ranged from

<table>
<thead>
<tr>
<th>Group</th>
<th>Leg</th>
<th>n</th>
<th>ART scores, % (mean ± SD)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Day 1</td>
<td>Day 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Day 3</td>
</tr>
<tr>
<td>Men</td>
<td>D</td>
<td>18</td>
<td>61.0±5.5</td>
<td>61.3±6.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>61.8±6.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ND</td>
<td>20</td>
<td>60.4±5.0</td>
<td>60.1±7.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.7±7.7</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>D</td>
<td>19</td>
<td>66.4±6.5</td>
<td>66.7±6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>66.7±7.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ND</td>
<td>19</td>
<td>67.1±7.2</td>
<td>67.4±7.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>67.3±7.5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>D</td>
<td>37</td>
<td>63.8±7.0</td>
<td>64.1±6.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64.3±7.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ND</td>
<td>39</td>
<td>63.9±7.7</td>
<td>63.8±8.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64.2±8.3</td>
<td></td>
</tr>
</tbody>
</table>

*Note: P-value is for repeated-measures analysis of variance of systematic bias between days 1, 2, and 3.*

**Abbreviations:** ART, Anteromedial Reach Test; D, dominant leg; ND, nondominant leg; n, number of participants included in analysis; SD, standard deviation.
0.93 to 0.97, while SEM values ranged from 1.25 to 1.77, demonstrating similar reliability and measurement error for both legs and sexes. For the total sample, the error associated with an individual’s score at a given point in time, at the 90% confidence level, was ±2.7% and the MDC90 was 3.8%.

**Interrater analysis**

Mean (± standard deviation) ART scores for raters A and B are shown in Table 4. There was no significant systematic bias between raters (Table 4). Reliability and measurement error statistics are presented in Table 5. ICC values ranged from 0.97 to 0.99, while SEM values ranged from 0.91 to 1.32, demonstrating similar reliability and measurement error for both legs and sexes.

**Discussion**

The ART demonstrated excellent intrarater and interrater reliability in both the dominant leg and nondominant leg of healthy men and women. ICC values exceeded 0.9, suggesting sufficient reliability for making clinical decisions regarding individuals.49 Such high reliability is not uncommon for ACL performance-based measures in healthy volunteers, who often demonstrate greater consistency than symptomatic patients.20 For example, intrarater reliability values exceeding 0.9 have been reported for hop and isokinetic testing in uninjured participants.20,31 Although interrater reliability of ACL performance-based measures has not been widely investigated, ICC values exceeding 0.9 have been reported for isokinetic testing in healthy volunteers.53

Only one previous study has evaluated reliability of the ART, and was presented in the form of a short conference abstract.24 As with our investigation, Rice et al demonstrated excellent intrarater reliability (ICC 0.96) of the ART in healthy volunteers;24 however, a flaw of this study is that reliability was calculated using reach distances that had not been normalized for leg length. Because non-normalized ART reach distances are related to leg length,25 they are effectively a surrogate measure of leg length and not a true indicator of an individual’s ability. As with our study, the ART scores of healthy participants should be normalized by expressing them as a percentage of reaching leg length.

The reliability of a measure similar to the ART, ie, the SEBT, has been evaluated in healthy volunteers by several authors.27,39,54,55 Two studies used normalized reach distances to calculate reliability and are comparable with

**Table 3 Intrarater reliability and measurement error statistics**

<table>
<thead>
<tr>
<th>Group</th>
<th>Leg</th>
<th>n</th>
<th>ICC (95% CI)</th>
<th>SEM, % (95% CI)</th>
<th>Error in an individual’s score, %</th>
<th>MDC90, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rater A</td>
<td>Rater B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>D</td>
<td>18</td>
<td>0.93 (0.86–0.97)</td>
<td>1.68 (1.35–2.19)</td>
<td>±2.8</td>
<td>3.9</td>
</tr>
<tr>
<td>ND</td>
<td>20</td>
<td>0.96 (0.92–0.98)</td>
<td>1.49 (1.22–1.92)</td>
<td>±2.4</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>D</td>
<td>19</td>
<td>0.97 (0.93–0.99)</td>
<td>1.25 (1.02–1.63)</td>
<td>±2.1</td>
<td>2.9</td>
</tr>
<tr>
<td>ND</td>
<td>19</td>
<td>0.94 (0.89–0.98)</td>
<td>1.77 (1.44–2.30)</td>
<td>±2.9</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>D</td>
<td>37</td>
<td>0.96 (0.93–0.98)</td>
<td>1.46 (1.26–1.75)</td>
<td>±2.4</td>
<td>3.4</td>
</tr>
<tr>
<td>ND</td>
<td>39</td>
<td>0.96 (0.93–0.97)</td>
<td>1.62 (1.40–1.93)</td>
<td>±2.7</td>
<td>3.8</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Error in an individual’s score estimated at 90% confidence level.

**Abbreviations:** D, dominant leg; ND, nondominant leg; n, number of participants included in analysis; ICC, intraclass correlation coefficient; 95% CI, 95% confidence interval; SEM, standard error of measurement; MDC90, minimum detectable change at 90% confidence level.

**Table 4 Mean Anteromedial Reach Test scores for raters A and B (interrater analysis)**

<table>
<thead>
<tr>
<th>Group</th>
<th>Leg</th>
<th>n</th>
<th>ART scores, % (mean ± SD)</th>
<th>P-value</th>
<th>Rater A</th>
<th>Rater B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>D</td>
<td>19</td>
<td>61.3±6.9</td>
<td>0.47</td>
<td>61.0±7.3</td>
<td>60.6±7.8</td>
</tr>
<tr>
<td>ND</td>
<td>19</td>
<td>60.5±7.8</td>
<td>60.2±7.6</td>
<td>0.33</td>
<td>60.2±7.6</td>
<td>60.2±7.6</td>
</tr>
<tr>
<td>Women</td>
<td>D</td>
<td>19</td>
<td>66.7±7.1</td>
<td>0.90</td>
<td>66.6±6.3</td>
<td>65.3±7.5</td>
</tr>
<tr>
<td>ND</td>
<td>19</td>
<td>65.3±7.5</td>
<td>68.0±6.7</td>
<td>0.10</td>
<td>67.3±7.5</td>
<td>68.0±6.7</td>
</tr>
<tr>
<td>Total</td>
<td>D</td>
<td>38</td>
<td>64.0±8.7</td>
<td>0.53</td>
<td>63.8±8.7</td>
<td>64.1±8.2</td>
</tr>
<tr>
<td>ND</td>
<td>38</td>
<td>63.9±8.3</td>
<td>63.1±8.2</td>
<td>0.50</td>
<td>63.9±8.3</td>
<td>64.1±8.2</td>
</tr>
</tbody>
</table>

**Note:** P-value is for paired t-test analysis of systematic bias between raters.

**Abbreviations:** ART, Anteromedial Reach Test; D, dominant leg; ND, nondominant leg; n, number of participants included in analysis; SD, standard deviation.

**Table 5 Interrater reliability and measurement error statistics**

<table>
<thead>
<tr>
<th>Group</th>
<th>Leg</th>
<th>n</th>
<th>ICC (95% CI)</th>
<th>SEM, % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rater A</td>
<td>Rater B</td>
</tr>
<tr>
<td>Men</td>
<td>D</td>
<td>19</td>
<td>0.97 (0.92–0.99)</td>
<td>1.32 (1.00–1.95)</td>
</tr>
<tr>
<td>ND</td>
<td>19</td>
<td>0.96 (0.91–1.00)</td>
<td>0.91 (0.69–1.33)</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>D</td>
<td>19</td>
<td>0.97 (0.92–0.99)</td>
<td>1.21 (0.92–1.80)</td>
</tr>
<tr>
<td>ND</td>
<td>19</td>
<td>0.98 (0.93–0.99)</td>
<td>1.09 (0.82–1.61)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>D</td>
<td>38</td>
<td>0.97 (0.95–0.99)</td>
<td>1.26 (1.02–1.62)</td>
</tr>
<tr>
<td>ND</td>
<td>38</td>
<td>0.98 (0.97–0.99)</td>
<td>1.04 (0.85–1.35)</td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** D, dominant leg; ND, nondominant leg; n, number of participants included in analysis; ICC, intraclass correlation coefficient; 95% CI, 95% confidence interval; SEM, standard error of measurement.
our investigation.\textsuperscript{39,55} Plisky et al found that intrarater reliability was good (range 0.85–0.88) for the three reach directions evaluated (anterior, posteromedial, posterolateral) but did not reach 0.9.\textsuperscript{55} Comparison with our investigation is limited, given that the SEBTam was not evaluated. Additionally, all measurements were taken during one test occasion, rather than several days apart, possibly reducing variation in performance. More recently, Munro and Herrington evaluated the intrarater reliability of all eight SEBT reach directions over 2 weeks.\textsuperscript{39} Reliability reached 0.9 for three of the eight reach directions (range 0.84–0.92), with an ICC of 0.85 for the SEBTam. Additionally, the MDC\textsubscript{90} for the SEBTam was 5.1%, which is higher than the 3.8% for the ART in our study.

The interrater reliability of the SEBT was evaluated in the aforementioned study by Plisky et al, with ICCs all exceeding 0.9 (range 0.99–1.00).\textsuperscript{39} Comparison with our investigation is again difficult, because the different raters simultaneously measured the same bout of the SEBT, rather than independently testing participants, possibly reducing performance variation. A shared finding with our investigation is that interrater reliability was superior to intrarater reliability. Given that our interrater investigation took place over approximately 30 minutes, it is not surprising that participants demonstrated less variability between bouts of the ART than for the intrarater investigation, which took place over several days.

**Clinical relevance**

Our study contains a number of clinically relevant findings:

1. There was no significant learning effect (systematic bias) between test occasions; therefore, no additional familiarization day is required before using the ART.
2. The error in an individual’s score at a given point in time was ±2.7% at the 90% confidence level. Therefore, if, for example, an individual is observed to score 60% on the ART, we can be 90% confident that they have scored at least 57.3% (ie, 60%–2.7%) and not more than 62.7% (ie, 60%+2.7%).\textsuperscript{29}
3. The MDC\textsubscript{90} was 3.8%. This is the smallest change in an individual’s score considered to be true change and not measurement error.\textsuperscript{29} Therefore, if an individual’s ART score improves or worsens between repeated tests by less than 3.8%, we can be 90% confident that they are unchanged.\textsuperscript{29} It should be noted that this value is for use with individuals only. For groups, the MDC\textsubscript{90} should be divided by $\sqrt{n}$.\textsuperscript{29} For example, in a group of 40 participants, the MDC\textsubscript{90} would be 0.6%.
4. The excellent interrater reliability suggests that a clinician who has not used the ART before can become proficient with the measure following a single familiarization session.
5. The excellent interrater reliability demonstrates that a 5-minute rest period can be used between the practice and recorded trials. Unlike the SEBT,\textsuperscript{29} the ART does not currently allow such a rest period.\textsuperscript{25} Although our data do not suggest any physical fatigue, some participants indicated that during days 2 and 3, when already familiarized with the ART, eight practice trials felt onerous and fatiguing. We will consider using a 5-minute rest period following the practice trials in future studies.

**Study limitations**

Our study was designed in accordance with recommendations for conducting a reliability study, considering such factors as sample size, blinding, representativeness of raters, systematic bias, appropriate statistical analysis, and clinical relevance.\textsuperscript{29,36,35,38,47} However, there are two main limitations that should be considered when interpreting its results. First, to exclude the effects of motor learning, the interrater reliability investigation did not take place until day 3. This meant that participants had already received instructions from rater A, ie, the more experienced rater. Therefore, any variability resulting from the initial instructions being given by different raters was removed, possibly inflating interrater reliability. However, given that the ART is simple to perform and instructions were standardized, any such inflation is likely to be small.

Second, although reported previously, outlier removal from a reliability study is not common practice and would have increased reliability. However, inspection of ART scores for the two excluded male participants supports the view that their results were anomalous. One excluded participant achieved consistent scores with rater A on days 1 and 2 and rater B on day 3, but then showed a decrease of 7.3% on the dominant leg with rater A on day 3. The fact that this participant had already achieved consistency over three test occasions with two different raters suggests that an error occurred during the final bout of the ART. The other excluded participant demonstrated an increase of 10.5% on the dominant leg between days 1 and 2. This resulted from scoring 14.7% less on the dominant leg than on the nondominant leg on day 1, but then attaining parity between legs for all subsequent test occasions. The mean
between-leg asymmetry for the rest of the sample on day 1 was 2.5%, suggesting that an error occurred during the first test occasion.

Table 6 shows the results of the six analyses from which outliers were originally excluded, with these two participants reincluded. The ICCs still exceed 0.9 in all but one case (intrarater reliability of the male dominant leg is now 0.89), demonstrating that reliability is not substantially affected. However, the effect on the measurement error statistics is more marked. For the total sample, the MDC90 for the dominant leg increases from 3.4% (Table 3) to 4.5% (Table 6), a factor increase of 1.3. For the male subgroup, the MDC90 for the dominant leg increases from 3.9% (Table 3) to 5.6% (Table 6), a factor increase of 1.4. Considering that the MDC90 for the male nondominant leg is 3.5%, this new value of 5.6% (from 3.9%) for the dominant leg seems abnormally high. This supports previous findings that just a small number of outliers can substantially inflate measurement error statistics.40 We believe that the inclusion of outliers in our analyses would have resulted in an unacceptable distortion of clinically meaningful values such as the MDC90, justifying the exclusion of these participants.

**Conclusion**

The ART demonstrated excellent levels of intrarater and interrater reliability in healthy volunteers, with no significant between-session learning effect. Reliability and measurement error were similar for both sexes (men and women) and legs (dominant and nondominant). The MDC90 was 3.8% and the error in an individual’s score at a given point in time was ±2.7%. Now that reliability of the ART has been demonstrated in healthy volunteers, future studies can investigate its measurement properties in ACL-injured patients.

**Disclosure**

The authors report no conflicts of interest in this work.

---

**Table 6 Reliability and measurement error statistics with outliers included**

<table>
<thead>
<tr>
<th>Group</th>
<th>Leg</th>
<th>Intrarater analysis</th>
<th>Interrater analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ICC (95% CI)</td>
<td>SEM, % (95% CI)</td>
</tr>
<tr>
<td>Men</td>
<td>D</td>
<td>0.89 (0.78–0.95)</td>
<td>2.42 (1.98–3.12)</td>
</tr>
<tr>
<td></td>
<td>ND</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>D</td>
<td>0.93 (0.89–0.96)</td>
<td>1.93 (1.67–2.29)</td>
</tr>
<tr>
<td></td>
<td>ND</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Note:** Error in an individual’s score estimated at 90% confidence level.

**Abbreviations:** D, dominant leg; ND, nondominant leg; ICC, intraclass correlation coefficient; 95% CI, 95% confidence interval; SEM, standard error of measurement; MDC90, minimum detectable change at 90% confidence level.

---

**References**


50. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34–42.


52. 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;57:1008–1018.

