Reliability and Validity of the Fullerton Functional Fitness Test: An Independent Replication Study

Jacqueline M. Miotto, Wojtek J. Chodzko-Zajko, Jennifer L. Reich, and Melissa M. Supler

A limiting factor in evaluating the functional status of older people is the lack of appropriate measurement tools for assessing functional mobility, muscle strength, aerobic endurance, agility, and flexibility. In this study, the reliability and validity of the seven-item Fullerton Functional Fitness Test (FFT) battery, designed for use with community-dwelling older adults, was examined. The test items were as follows: floor sit-and-reach, back scratch, 8-ft up-and-go, arm curl, 30-s chair stand, 2-min step, and 9-min walk. Seventy-nine participants (42 physically active, 37 sedentary) completed the FFT battery three times within a 2-week period. The test–retest reliability intraclass correlation coefficients were high. Construct validity analysis revealed that five of the seven FFT items discriminated between the physically active and sedentary groups. In conclusion, most of the evidence from the stability reliability and discriminant validity analyses supports the view that the Fullerton FFT battery is a reliable and valid test of functional fitness.

Key Words: aging, functional capacity, reliability, validity, assessment

Spirduso (1995) has proposed that the older adult population can be categorized along a five-point continuum of physical functioning, with physically fit seniors at one end of the continuum and physically frail, dependent seniors at the other. A recent report by the World Health Organization (1997) suggests that the benefits of regular physical activity are not restricted to any particular group along this continuum but that significant benefits can be gained from physical activity by almost all older adults, regardless of their health status and functional ability. Unfortunately, the factors that influence transitions between one point on the continuum and another are poorly understood (Chodzko-Zajko, 1996). For example, little is known about the factors that precipitate an individual’s transition from functional independence to physical frailty. Some sedentary people can

J.M. Miotto is with the Department of Physical Therapy, Walsh University, North Canton, OH 44720. W.J. Chodzko-Zajko is with the School of Exercise, Leisure, and Sport, Kent State University, Kent, OH 44242-0001. J.L. Reich is with the Benjamin Rose Institute, Cleveland, OH 44114. M.M. Supler is with the Physical Therapy Department, Cleveland State University, Cleveland, OH 44115.
maintain an independent lifestyle for many decades, whereas others regress far more rapidly into frailty and dependency. Several gerontologists have suggested that our understanding of the temporal dynamics of senescence is unlikely to be improved until more sensitive measures of functional fitness are developed (Rikli & Jones, 1997; Chodzko-Zajko).

Chodzko-Zajko (1996) and Rikli and Jones (1997) have argued that the absence of functional fitness measures that are sufficiently sensitive to assess functional status across a large portion of the physical-functioning continuum is the single largest measurement issue in exercise gerontology. It is generally accepted that such a fitness battery must assess multiple components of physical functioning. Recently, researchers at the Ruby Gerontology Center at California State University-Fullerton developed a Functional Fitness Test (FFT) battery consisting of seven items, designed to permit the evaluation of functional fitness across the physical-functioning continuum (Rikli & Jones, 1999a). The Fullerton FFT items are designed to evaluate the following components of functional fitness: aerobic endurance, upper and lower body muscle strength and endurance, upper and lower body flexibility, and agility and dynamic balance. The purpose of this study was to perform an independent assessment of the test–retest reliability and construct validity of the Fullerton FFT battery in 79 older adults. Rikli and Jones (1999a) have examined the test–retest reliability (over two test sessions) and construct validity in their laboratory, but, considering that the FFT is to be used nationwide and that national norms are available (Rikli & Jones, 1999b), it is important to verify the test–retest reliability and construct validity of the test items in different settings.

**Methods**

**PARTICIPANTS**

Seventy-nine older adults volunteered to participate in the study. Forty-two were physically active individuals (PA) who had participated in regular physical activity at least three times a week for the preceding 8 months. The 37 participants who were sedentary (SED) were recruited from local senior citizen groups and retirement centers in the community. These individuals reported no participation in structured exercise for the last 3 years. The PA and SED groups were similar in age range (60–86 years) and sex distribution (see Table 1). All participants were living independently in the community.

Participants gave informed consent and obtained the informed consent of their physicians prior to taking part in the study. All participants completed a general health questionnaire and the Physical Activity Readiness Questionnaire (Reading & Shephard, 1992) to screen for cardiovascular, orthopedic, or other medical contraindications to physical activity.

**PROCEDURES**

Because one of the goals of this study was to examine the test–retest reliability of the FFT items, the battery was performed in its entirety three times within a 2-week
Table 1  Descriptive Data by Group (PA vs. SED) and Sex

<table>
<thead>
<tr>
<th>Variable</th>
<th>Physically Active</th>
<th>Sedentary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men, n = 24</td>
<td>Women, n = 18</td>
</tr>
<tr>
<td>Age</td>
<td>67.75</td>
<td>68.33</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>173.18</td>
<td>160.13</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>80.23</td>
<td>64.49</td>
</tr>
<tr>
<td>BMI</td>
<td>26.71</td>
<td>25.21</td>
</tr>
<tr>
<td>SBP</td>
<td>128.00</td>
<td>127.18</td>
</tr>
<tr>
<td>DBP</td>
<td>75.91</td>
<td>77.06</td>
</tr>
</tbody>
</table>

period. In order to better reflect clinical reality—the same individual is not always available to administer tests performed on different days—we did not require that the same tester administer repeated tests on the same participant; rather, all tests were conducted by a team of similarly trained undergraduate and graduate students, with no specific attention paid to the assignment of participants to testers. Accordingly, our reliability measures include a component of interrater variability.

It is important to note that after the completion of the present study, two modifications were made to the Fullerton FFT battery (Rikli & Jones, 1999a). Specifically, the protocols for two of the test items have been changed. The 9-min walk has been modified to a 6-min walk, and the floor sit-and-reach has been changed to a chair sit-and-reach. Data reported in this study reflect the original versions of these two test items. The complete protocols for all test items are described by Rikli and Jones (1999a). The protocols for the original 9-min walk and floor sit-and-reach are described below in their entirety. Additionally, any deviations from the Rikli and Jones protocols are described below for each test item.

All participants engaged in 8 min of supervised warm-up and stretching exercises before testing.

**Floor Sit-and-Reach.** This test was designed to assess lower body, primarily hamstring, flexibility. The original Fullerton FFT item was a modification of the sit-and-reach test from Osness et al. (1987). The heels of the participant are placed at the 20-in. mark for the original FFT item, instead of the 25-in. mark as was done by Osness et al. For the FFT, each participant sat on a platform, legs extended, with a yardstick positioned between the legs, with the zero end of the yardstick proximal. The participant’s heels were positioned at the 20-in. mark, 12 in. apart. The tester ensured that the heels remained in place with the toes pointing up and the legs straight. The participant slowly bent forward, sliding his or her hands along the yardstick as far as possible. The final position was held for 2 s. Four trials were performed, and the scores were measured to the nearest 1/2 in. All four trials were recorded, and the best performance was used as the score.
Back Scratch. This test was designed to assess upper body and shoulder flexibility (Rikli & Jones, 1999a). The protocol described by Rikli and Jones was followed, with the exception that all four trials were recorded (instead of only the last two), and the best of the four trials was used as the score (instead of the better of the last two trials).

8-Foot Up-and-Go. This test was designed to measure physical agility and dynamic balance (Rikli & Jones, 1999a). The score was the time it took to complete the agility course. The protocol described by Rikli and Jones was used except that three test trials, instead of one practice trial and two test trials, were given, and the score recorded was the best of the three performances.

Arm Curl. The arm curl test was designed to measure arm (biceps) muscle strength and endurance (Rikli & Jones, 1999a). The score was the total number of curls made correctly within the 30-s time limit. The Rikli and Jones protocol was followed, except that three trials were performed instead of one. One-minute rest periods were given between trials. For this study, the best of the three trials was the score.

30-Second Chair Stand. This item was designed to assess lower body muscle strength and endurance (Rikli & Jones, 1999a). The protocol described by Rikli and Jones was followed, and the score was the total number of stands executed within 30 s.

2-Minute Step Test. This test was designed to measure general aerobic endurance and lower body muscle endurance (Rikli & Jones, 1999a). The test score was the number of right-leg steps taken in 2 min. The Rikli and Jones protocol was followed.

9-Minute Walk. This test was designed to measure aerobic endurance (Rikli & Jones, 1998). The protocol for the 9-min walk is similar to the new, revised 6-min walk test described by Rikli and Jones (1999a), except that the original version is 3 min longer. The participants were instructed to walk around a measured course for 9 min and to try to cover as much distance as possible. They were told to pace themselves—that they should not end the test completely exhausted or unduly fatigued. To assist with pacing, the participants were told when 5 min, 7 min, and 8 min had elapsed. The score was the total number of yards walked in 9 min, to the nearest 5 yd.

LABORATORY MEASURES

In addition to the FFT measures described above, several other physiologic and anthropometric measures were assessed. Systolic blood pressure (SBP), diastolic blood pressure (DBP), resting heart rate (RHR), height, weight, and body mass index (BMI) were measured for each participant. Height and weight were recorded according to standard experimental procedure, using a calibrated physician’s scale and stadiometer. BMI was calculated as height (in meters) squared divided by weight (in kilograms). Participants were requested to recline on a bed for 10 min, after which RHR and resting SBP and DBP were recorded.

STATISTICAL ANALYSIS

Analysis was performed using SPSS software. The intraclass correlation coefficient (ICC), R, from a one-way repeated-measures ANOVA (Baumgartner, 1989) was
used to estimate the test–retest reliability of the FFT items. Two ICCs were calculated. First, a multisession ICC was calculated across the stable test sessions using Formula 1:

\[ R = \frac{MS_S - MS_w}{MS_S} \]  \hspace{1cm} (1)

Second, a single-session ICC was estimated using Formula 2:

\[ R = \frac{MS_S - MS_w}{MS_S + \left(\frac{K}{K'} - 1\right)(MS_w)} \]  \hspace{1cm} (2)

The single-session ICC reflects what the reliability would have been if only one test session had been given (Safrit & Wood, 1995). This should be of interest to those who only expect to be administering the test on one occasion. In Formulas 1 and 2, \( R \) is the ICC, \( MS_S \) is the mean square between, \( MS_w \) is the mean square within, \( K \) is the number of repeated measures administered, and \( K' \) is the number of repeated measures for which \( R \) is estimated (Safrit & Wood).

The ICC reflects the relative reliability of the test scores, or the extent to which the scores are in the same relative position from one test session to another (Baumgartner, 1989). Absolute reliability is a measure of scoring consistency that shows the degree to which the individual’s scores do not change (Baumgartner). Absolute reliability was evaluated across the three test sessions using a one-way repeated-measures ANOVA followed by post hoc multiple comparison tests. Specifically, a Tukey’s HSD (\( p \leq .05 \)) test was used to determine which sessions were different from one another. The best score from each of the three test sessions was used to evaluate both relative and absolute reliability.

The “known-groups,” or “group-differences,” method (Morrow, Jackson, Disch, & Mood, 1995) was used to assess construct, or discriminant, validity. Specifically, the construct validity of the FFT items was evaluated by assessing their ability to discriminate between discrete groups of physically active (PA) and sedentary (SED) individuals. For this analysis a two-way group (PA vs. SED) by sex (male vs. female) ANOVA was used. Bonferroni’s correction was employed to adjust for the seven multiple comparisons, whereby the \( p \) value of .05 was divided by 7 and statistical significance was set at \( p \leq .007 \).

Results

DESCRIPTIVE DATA

Descriptive data for the physically active and sedentary groups are presented in Table 1. Group and sex differences for the descriptive variables were examined using a two-way group-by-sex ANOVA, which revealed no significant group-by-sex interactions. Groups did not differ with respect to age or height, but participants in the PA group weighed less, \( F(1, 74) = 5.4, p = .021 \), had significantly lower BMIs, \( F(1, 74) = 5.6, p = .020 \), and had significantly lower systolic blood pressures, \( F(1, 70) = 3.9, p = .050 \). There were no group differences in DBP.
With respect to sex comparisons, men, as expected, were heavier, $F(1, 74) = 39.6, p < .001$, and taller, $F(1, 74) = 60.8, p < .001$. The BMI for men was also significantly higher than for women, $F(1, 74) = 6.3, p = .014$. No differences were found between men and women for age, SBP, or DBP.

**TEST–RETEST RELIABILITY**

ICCs and between-sessions performance differences are presented in Table 2. Means and standard deviations for the FFT items across all three test sessions are presented in Table 3. For the floor sit-and-reach, the one-way ANOVA showed significant differences across sessions, $F(2, 154) = 6.4, p = .002$. Post hoc analysis revealed that sit-and-reach performance did not differ significantly from Session 1 to Session 2 but that performance did improve between Sessions 2 and 3. The ICC, calculated across the first and second test sessions, was found to be high, $R = .95$. The estimated single-session ICC was also high, $R = .92$.

Results for the back scratch showed no differences in test performance between test sessions, $F(2, 156) = 1.3, p = .267$. The ICC across all three test sessions was found to be high, $R = .98$, as was the single-session ICC, $R = .94$.

There were significant differences across sessions for the 8-ft up-and-go test, $F(2, 156) = 15.2, p < .001$. Post hoc comparisons revealed that 8-ft up-and-go performance did not differ from Session 1 to Session 2, but performance improved significantly from Session 2 to Session 3. The test–retest reliability ICC calculated across the stable test sessions (i.e., Sessions 1 and 2) was found to be high, $R = .94$. The ICC for a single session was acceptable, $R = .86$.

For the arm curl test, results of the one-way ANOVA revealed significant differences across sessions, $F(2, 156) = 15.4, p < .001$. It was found from post hoc tests that arm curl performance improved significantly from Session 1 to Session 2 but not from Session 2 to Session 3. The ICC across Sessions 2 and 3 was found to be high, $R = .96$, and the single-session ICC was acceptable, $R = .88$.

<table>
<thead>
<tr>
<th>FFT item</th>
<th>ICC (R)</th>
<th>Sessions*</th>
<th>ICC (R, single-session)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor sit-and-reach</td>
<td>.95</td>
<td>1 2 3</td>
<td>.92</td>
</tr>
<tr>
<td>Back scratch</td>
<td>.98</td>
<td>1 2 3</td>
<td>.94</td>
</tr>
<tr>
<td>8-ft up-and-go</td>
<td>.94</td>
<td>1 2 3</td>
<td>.86</td>
</tr>
<tr>
<td>Arm curl</td>
<td>.96</td>
<td>1 2 3</td>
<td>.88</td>
</tr>
<tr>
<td>30-s chair stand</td>
<td>.96</td>
<td>1 2 3</td>
<td>.90</td>
</tr>
<tr>
<td>2-min step</td>
<td>.95</td>
<td>1 2 3</td>
<td>.83</td>
</tr>
<tr>
<td>9-min walk</td>
<td>.97</td>
<td>1 2 3</td>
<td>.91</td>
</tr>
</tbody>
</table>

*The sessions underlined are not significantly different at $p \leq .05$. 
Table 3  *Ms* and *SDs* for the FFT Items in Each of the 3 Test Sessions

<table>
<thead>
<tr>
<th>FFT item</th>
<th>Session 1, n = 79</th>
<th></th>
<th>Session 2, n = 79</th>
<th></th>
<th>Session 3, n = 79</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M</em></td>
<td><em>SD</em></td>
<td><em>M</em></td>
<td><em>SD</em></td>
<td><em>M</em></td>
<td><em>SD</em></td>
</tr>
<tr>
<td>Floor sit-and-reach (in.)</td>
<td>15.19</td>
<td>4.58</td>
<td>15.69*</td>
<td>4.76</td>
<td>16.00*</td>
<td>4.65</td>
</tr>
<tr>
<td>8-ft up-and-go (s)</td>
<td>5.69*</td>
<td>1.24</td>
<td>5.57*</td>
<td>1.19</td>
<td>5.31</td>
<td>1.07</td>
</tr>
<tr>
<td>Arm curl (no. in 30 s)</td>
<td>19.76</td>
<td>4.52</td>
<td>20.66*</td>
<td>5.03</td>
<td>21.20*</td>
<td>4.93</td>
</tr>
<tr>
<td>30-s chair stand (no.)</td>
<td>14.24</td>
<td>3.69</td>
<td>14.76</td>
<td>3.95</td>
<td>15.33</td>
<td>3.91</td>
</tr>
<tr>
<td>2-min step (no.)</td>
<td>104.63*</td>
<td>25.15</td>
<td>108.00*</td>
<td>26.21</td>
<td>110.65*</td>
<td>23.73</td>
</tr>
<tr>
<td>9-min walk (yd)</td>
<td>958.2</td>
<td>148.7</td>
<td>978.6*</td>
<td>154.0</td>
<td>984.0*</td>
<td>146.4</td>
</tr>
</tbody>
</table>

*The means with asterisks are not significantly different at *p* ≤ .05.

Results of the 30-s chair stand revealed significant differences across sessions, *F*(2, 156) = 15.2, *p* = .001, and post hoc tests showed that performance improved significantly across all three test sessions. Both the ICC for test–retest reliability across the three test sessions, *R* = .96, and the estimated ICC for a single session, *R* = .90, were found to be high. Because a stable score was not obtained across the three test sessions, however, caution is warranted when interpreting the ICC (Baumgartner, 1989).

There were significant differences across sessions for the 2-min step test, *F*(2, 156) = 8.2, *p* = .001. Post hoc tests showed that test performance improved significantly from Session 1 to Session 3 but did not differ significantly from Session 1 to Session 2 or from Session 2 to Session 3. The ICC calculated across all three sessions was high, *R* = .95, and the single-session ICC was acceptable, *R* = .83.

Finally, for the 9-min walk, there were significant differences across sessions, *F*(2, 156) = 8.6, *p* = .001. Post hoc tests revealed that 9-min walk performance improved significantly from Session 1 to Session 2, but there was no significant improvement in distance walked from Session 2 to Session 3. The ICC for Sessions 2 and 3 and was found to be high, *R* = .97, as was the single-session ICC, *R* = .91.

**CONSTRUCT VALIDITY OF THE FFT BATTERY**

In order to examine the construct validity of the FFT items, the ability of the individual FFT items to discriminate between discrete groups was assessed using a two-way group-by-sex ANOVA. Because there were no significant group-by-sex interactions for any of the FFT items, Tables 4 and 5 report only the main effects of group and sex.

There were no differences in floor sit-and-reach between the PA and SED groups, but there was a significant main effect for sex, *F*(1, 74) = 34.3, *p* < .001, with the women (18.5 in. ± 3.5) reaching farther than the men did (13.6 in. ± 4.4).
Table 4  Functional Fitness Test Scores by Group (PA vs. SED)

<table>
<thead>
<tr>
<th>FFT item</th>
<th>Physically active, $n = 42$</th>
<th>Sedentary, $n = 37$</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor sit-and-reach (in.)</td>
<td>16.60</td>
<td>15.31</td>
<td>4.7</td>
<td>.032</td>
</tr>
<tr>
<td>Back scratch (in.)</td>
<td>-2.04</td>
<td>-2.77</td>
<td>1.5</td>
<td>.218</td>
</tr>
<tr>
<td>8-ft up-and-go (s)</td>
<td>4.96</td>
<td>5.71</td>
<td>9.3</td>
<td>.003</td>
</tr>
<tr>
<td>Arm curl (no. in 30 s)</td>
<td>22.95</td>
<td>19.22</td>
<td>11.8</td>
<td>.001</td>
</tr>
<tr>
<td>30-s chair stand (no.)</td>
<td>17.07</td>
<td>13.35</td>
<td>22.4</td>
<td>.001</td>
</tr>
<tr>
<td>2-min step (no.)</td>
<td>121.38</td>
<td>98.46</td>
<td>22.1</td>
<td>.001</td>
</tr>
<tr>
<td>9-min walk (yd)</td>
<td>1,030.6</td>
<td>931.05</td>
<td>9.1</td>
<td>.004</td>
</tr>
</tbody>
</table>

Table 5  Functional Fitness Test Scores by Sex (male vs. female)

<table>
<thead>
<tr>
<th>FFT item</th>
<th>Men, $n = 41$</th>
<th>Women, $n = 38$</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor sit-and-reach (in.)</td>
<td>13.61</td>
<td>18.51</td>
<td>34.3</td>
<td>.001</td>
</tr>
<tr>
<td>Back scratch (in.)</td>
<td>-4.45</td>
<td>-0.14</td>
<td>19.6</td>
<td>.001</td>
</tr>
<tr>
<td>8-ft up-and-go (s)</td>
<td>4.99</td>
<td>5.66</td>
<td>6.8</td>
<td>.011</td>
</tr>
<tr>
<td>Arm curl (no. in 30 s)</td>
<td>22.49</td>
<td>19.82</td>
<td>4.7</td>
<td>.032</td>
</tr>
<tr>
<td>30-s chair stand (no.)</td>
<td>15.71</td>
<td>14.92</td>
<td>.133</td>
<td>.716</td>
</tr>
<tr>
<td>2-min step (no.)</td>
<td>117.00</td>
<td>103.79</td>
<td>5.2</td>
<td>.026</td>
</tr>
<tr>
<td>9-min walk (yd)</td>
<td>1,044.9</td>
<td>918.2</td>
<td>16.3</td>
<td>.001</td>
</tr>
</tbody>
</table>

Similarly, the back scratch test revealed no main effect for group, but there was a significant main effect for sex, $F(1, 75) = 19.6, p < .001$, with the women ($-0.14$ in. ± 3.72) reaching about 4 in. farther than the men did ($-4.45$ in. ± 4.91).

The 8-ft up-and-go test revealed no main effect for sex but a significant main effect for group, $F(1, 75) = 9.3, p = .003$. The PA group was significantly faster (4.96 ± 1.02 s) than the SED group (5.71 ± 1.01 s). There was a significant difference between groups for the arm curl test, $F(1, 75) = 11.8, p < .001$. The PA group completed significantly more arm curls (22.95 curls ± 5.11) than the sedentary group did (19.22 curls ± 3.92). There was no main effect for sex. Likewise, for the 30-s chair-stand test, there was no main effect for sex, but there was a significant main effect for group, $F(1, 75) = 22.4, p < .001$. The PA group completed significantly more stands (17.07 stands ± 3.58) in the 30-s time period than the SED group did (13.35 stands ± 3.29).
There was a significant main effect for group, $F(1, 75) = 22.1, p < .001$, for the 2-min step test. The PA group completed significantly more steps in the 2-min time period (121.38 steps ± 20.45) than the SED group did (98.46 steps ± 21.38). There was no main effect for sex.

The 9-min walk test revealed significant main effects for group, $F(1, 75) = 9.1, p = .004$, and sex, $F(1, 75) = 16.3, p < .001$. In the 9-min time period, the PA group walked significantly farther (1,030.62 yd ± 147.75) than the SED group did (931.05 yd ± 127.03), and the men walked significantly farther (1,044.93 yd ± 137.60) than the women did (918.24 yd ± 127.03).

**Discussion**

One objective of this study was to examine the test–retest reliability of the FFT battery. In order to accurately assess an individual’s functional capacity, it is important to minimize the influence of extraneous factors such as practice or learning. To do this, it is necessary to determine the appropriate number of sessions required for each item within a test battery. This can be accomplished by establishing the point at which stable performance is achieved across multiple test sessions. Examination of the test–retest reliability revealed significant heterogeneity across tests with respect to the number of sessions required to achieve stable performance.

For the floor sit-and-reach test, performance was stable across Sessions 1 and 2 but improved from Session 2 to Session 3. If only one test session had been used, the floor sit-and-reach score would have underestimated the final maximum score by 3.92%. Based on these results, it is recommended that three test sessions (two practice, one test) be administered for the floor sit-and-reach test to achieve a stable score. For Sessions 1 and 2, it was observed that only 4% of the best scores were performed in Trial 1 or 2, and for Session 3, all of the best scores were performed in either the third or the fourth trial. Therefore, a protocol in which only the last two trials are recorded appears appropriate.

Bravo et al. (1994) examined the test–retest reliability of the floor sit-and-reach test across two testing days and found no difference in performance between the two days. Shaulis, Golding, and Tandy (1994) examined the test–retest reliability across three testing days and found a significant improvement from Session 1 to Session 3 but no performance difference between Sessions 1 and 2 or between Sessions 2 and 3. The reported ICCs for these studies, .94 and .97, respectively, were similar to those in the present study.

Unlike the floor sit-and-reach test, the best scores for the back scratch from each consecutive test session did not continue to show improvement. The ICC calculated across test sessions was high, .98, indicating a reliable measure across all of the three test sessions. These results indicate that only one test session is necessary to achieve a stable score. Rikli and Jones (1999a) also report a high ICC, $R = .96$, and no improvement in performance across two test sessions for the back scratch test. For all three test sessions, it was observed that 16–19% of the best scores were performed in Trial 1 or 2. Therefore, a protocol in which the score is the best of all four of the trials, instead of only the last two trials (Rikli & Jones, 1999a), might be more appropriate.

As with the floor sit-and-reach, the best scores from Sessions 1 and 2 for the 8-ft up-and-go were not significantly different, but there was significant improvement
in the best scores from Session 2 to Session 3. It is unclear why performance was stable across Sessions 1 and 2 but then improved with Session 3. If practice or learning effects contributed to improved performance from Session 2 to Session 3, why didn’t these same effects influence performance from the first to the second session? One possible explanation is that for the last test session, the participants put forth a greater effort because they were more comfortable with the protocol and testing situation, and so felt more confident that they would not slip or lose balance when turning or rising quickly from the chair. Another possibility is that over the first two sessions the participants might have developed strategies that resulted in faster performance in the last session. From these results, it is suggested that three test sessions be completed to minimize any effects from practice or learning and to reach a stable score.

Unlike the Rikli and Jones (1999a) protocol (one practice trial and two test trials), three test trials were performed for the 8-ft up-and-go test in this study. It was observed that across the test sessions the best score was from the first trial 11–20% of the time, from the second trial 20–25% of the time, and from the third trial 55–68% of the time. These results indicate that at least three consecutive trials can be performed without deleterious fatigue effects influencing performance, and that perhaps three trials are more appropriate than two.

Rikli and Jones (1999a) reported a high ICC and no improvement in test performance across two test sessions for the 8-ft up-and-go test. Additionally, previous research has reported reliability results on an agility test similar to the 8-ft up-and-go. Bravo et al. (1994) and Shaulis et al. (1994) examined the test–retest reliability for a similar agility task, and both groups found a high ICC, but, as with the present study, neither attained a stable score. Instead, test performance improved from session to session. The lack of stability across test sessions in these two studies was also considered to be the result of practice or learning effects.

For the arm curl test, there was significant improvement in test performance from Session 1 to Session 2 but no improvement from Session 2 to Session 3. If only one session had been completed, the score would have been underestimated by 7.30%. These results suggest that two test sessions should be administered. It was also observed that the best score was from the third trial 70–72% of the time for all three test sessions. These data indicate that fatigue does not adversely influence performance from trial to trial when three trials are conducted.

As with this study, Bravo et al. (1994), Shaulis et al. (1994), and Rikli and Jones (1999a) all showed a high ICC over multiple test sessions for the arm curl test. Bravo et al. examined performance over two test sessions, and a stable score was not achieved. Unlike the present study, Shaulis et al.'s work revealed continued improvements in test performance across three successive test sessions. Rikli and Jones, however, showed no improvement in performance across two test sessions.

Although the high ICC for the 30-s chair stand test showed good relative reliability, there was significant improvement in test performance across all three test sessions. Once again, practice or learning effects might explain the continued improvement in test performance across sessions. It is therefore recommended that at least three test sessions be completed for the 30-s chair stand test to control for learning or practice effects. Additional research might determine the point at which a stable score can be reached. Because of potential fatigue effects, this test was administered only once per test session. Increasing the number of trials per session
and ensuring adequate rest between trials might increase the possibility of achieving a stable score in a reasonable number of sessions. Also, increasing the amount of practice or level of motivation might help stabilize performance. Rikli and Jones (1999a) reported stable performance across the two test sessions in their reliability study.

Performance on the 2-min step test was stable across all three test sessions, and the ICC showed high relative reliability. Although Rikli and Jones (1999a) recommend one practice and one test trial, based on the present study data, only one trial appears necessary for stable performance. Rikli and Jones (1999a) found acceptable relative and absolute reliability for the 2-min step test across two sessions that followed a practice session.

Participants reached stable performance on the 9-min walk test by Session 2, and the ICC for Sessions 2 and 3 displayed high relative reliability. From these results, it is recommended that two test sessions be administered for the 9-min walk. Prior exposure to this test probably enables participants to pace themselves better throughout the 9-min walking period. Stable performance was also obtained within two sessions, after a practice session, by Rikli and Jones (1999a).

Previous research has examined the reliability of a similar walk test—a timed 880-yd walk. Bravo et al. (1994) administered two test sessions and did not achieve a stable score. Shaulis et al. (1994) examined the reliability of the 880-yd walk over three test sessions and found, similar to the present study, improved performance from Session 1 to Session 2 but no improvement from Session 2 to Session 3. Both of these groups found acceptable relative reliability.

A secondary goal of the investigation was to determine the construct, or discriminant, validity of the FFT items. The known-groups method (Morrow et al., 1995) was used to examine the ability of the individual test items to discriminate between the PA and SED groups. It was found that five of the seven FFT items successfully discriminated between the two groups, which provides support for the construct validity of the FFT.

The two flexibility items did not discriminate between the PA and SED groups. This might not be surprising, because although flexibility is considered an important component of fitness, it is not always related to performance on tests of aerobic fitness and muscle strength and endurance, and active people do not always have better flexibility than sedentary individuals do (Maud & Cortez-Cooper, 1995). Unlike in this study, Rikli and Jones (1999a) showed differences in flexibility between active and sedentary older adults on the back scratch and chair sit-and-reach tests.

Examination of the floor sit-and-reach test scores from this study suggest they are comparable to those reported in previously published research. Floor sit-and-reach performance averaged 18.5 in. for the women and 13.6 in. for the men. In general, and as is shown in this study, women perform better than men do on flexibility items. For example, Shaulis et al. (1994) reported significant gender differences between women (20.08 in.) and men (12.26 in.) on the floor sit-and-reach test. Osness et al. (1996) have published norms for the floor sit-and-reach test for 5-year age ranges from 60 to 90. Their average score from age 60 to age 90 for men is 15.9–11.2 in., and for women it is 18.2–14.5 in. In a sample of Korean women, Kim and Tanaka (1995) reported an average floor sit-and-reach of 18.03 in. Therefore, although the scores on the floor sit-and-reach test did not differ
between the PA and SED groups, they were similar to other reported floor sit-and-reach scores. (The values for the sit-and-reach test from Shaulis et al., Osness et al., and Kim and Tanaka have been adjusted because of the different heel placement—25-in. mark instead of 20-in. mark—so as to be comparable to the present study.)

The upper body flexibility item did not discriminate between the PA and SED groups, either, but the scores from this test are similar to those in other studies. In this study, the average score on the back scratch test was −0.14 in. for women and −4.45 in. for men. Rikli and Jones (1999b) reported average back scratch performance of −1.9 in. for women and −4.8 in. for men ages 60–94.

Significant differences between the PA and SED groups were detected in the 8-ft up-and-go. The PA participants averaged a score of 4.96 s, and the SED participants averaged 5.71 s—about 13% slower. Rikli and Jones (1999a) reported an average score of 6.0 s for exercisers and 7.1 s for nonexercisers in the 8-ft up-and-go test. They also reported average performance ranging from 5.2 to 7.1 s for men and women 60 to 89 years old. The faster performance on the 8-ft up-and-go in the current study, compared with the Rikli and Jones (1999a) study, might partially reflect a more fit and mobile study population, but it also might be a result of age. The average age of the exercisers in Rikli and Jones’s study was 75.6 years, and of nonexercisers 77.7 years, compared with 68 and 69 years for the PA and SED groups, respectively, in this study.

Both of the muscle strength and endurance test items, the arm curl and the 30-s chair stand, discriminated successfully between the PA and SED groups. There were no sex differences in performance for either of these two tests. The PA group averaged 23.0 arm curls, and the SED group averaged only 19.2 curls. Osness et al. (1996) reported average scores for different age groups: The average scores ranged from 23.7 arm curls for 60- to 64-year-old men to 17.8 for 85- to 90-year-old men. For women, average scores ranged from 21.8 arm curls for 60- to 64-year-old women to 16.2 for 85- to 90-year-old women. Rikli and Jones (1999a) reported an average score of 19.8 to 16.5 curls on the arm curl test for men and women 60 to 89 years old and that exercisers (18.7 curls) performed significantly better on the arm curl than nonexercisers did (15.5 curls).

With regard to the 30-s chair stand test, the average score for the PA group was 17.1 stands and for the SED group was 13.4 stands. Rikli and Jones (1999a) also reported differences between exercisers (13.3 stands) and nonexercisers (10.8 stands). Average test performance for 60- to 89-year-old men and women ranged from 14.0 to 11.9 stands. Again, the higher scores reported on both the 30-s chair stand and the arm curl in the present study, compared with Rikli and Jones (1999a), are likely caused by differences in fitness and age between the two study groups.

The two aerobic endurance items, the 2-min step and the 9-min walk, also successfully discriminated between the PA and SED groups. The average number of steps taken in 2 min for the PA group was 121.4 and for the SED group was 98.5. There were no sex differences in performance. Rikli and Jones (1999a) have reported average test scores ranging from 100.4 to 83.5 steps in 60- to 89-year-old men and women. They also reported differences in performance between exercisers (95.8 steps) and nonexercisers (72.8 steps). It was observed that performance on the 2-min step test in the current study was better than performance reported by Rikli and Jones (1999a). Once again, this observed performance difference might result from differences in the fitness level and age of the individuals in the two studies.
In the 9-min walk test, the PA group walked an average of 1,030.6 yd, and the SED group walked an average of 931.1 yd. There was also a sex difference in performance, with the men walking 13.8% farther than the women did. Because the 9-min walk test has not been used in previously published research, it is more logical to compare speeds of walking between studies. In the present study, the average walk speed was 3.73 mph (1.67 m/s) overall—3.96 mph (1.77 m/s) for men and 3.48 mph (1.55 m/s) for women. In an 880-yd walk test, Bazzano, Cunningham, Cama, and Falconio (1995) found that healthy older adult men and women walked at 3.92 mph (1.75 m/s). Shaulis et al. (1994) reported walking speeds of 3.66 mph (1.64 m/s) for healthy women and 4.08 mph (1.82 m/s) for healthy men for the same 880-yd walk test. Finally, Rikli and Jones (1999a) found that their physically active older men and women walked at an average speed of 3.68 mph (1.64 m/s).

It is important to recognize that in the present investigation, all of the participants were volunteers. The reliance on volunteer participants creates a selection bias and distinguishes the sample from the overall older adult population (Chodzko-Zajko, 1996). In addition, the mode of advertising the research study—local newspapers, senior center flyers, retirement community notices—could also have created a selection bias toward people who receive and read newspapers and those who live in or visit certain locales in the community. This advertising strategy might have resulted in a selection bias toward individuals of middle to higher socioeconomic status. Likewise, the requirement of a physician’s consent for participation could also bias the sample selection toward older adults who have the means to visit a doctor and those who are willing to schedule such a visit.

The majority of the PA group was recruited from active participants in a university-based senior exercise program. This convenient sample of physically active individuals might have been more highly motivated to participate in a research study than the broader population of physically active older adults, and therefore the sample might not be representative of the active older adult population as a whole. Nonetheless, the mode of recruitment and use of physician consent in this study is not dissimilar to other studies in exercise and aging research. For example, Mobily and Mobily (1997) recruited participants from an ongoing fitness study to examine the reliability of a fitness test battery, and Bravo et al. (1994) recruited individuals from both an exercise class and an ongoing bone mass study for their reliability and validity study.

**Conclusions and Recommendations**

In summary, although the relative reliability for all of the test items was acceptable, the results of the absolute reliability analysis indicated a learning or practice effect on most of the test items. Consequently, the number of test sessions for the FFT items is of great importance, particularly when using scores as baseline data for an intervention study. It is also important to keep recommendations feasible and practical. Therefore, if the entire test battery is to be used and time is an important factor, it is recommended that one practice session be completed before the official test session. However, this protocol might lead to an underestimation of the maximal performance on the floor sit-and-reach, chair stand, and 8-ft up-and-go. If determining an individual’s maximal performance is the goal, three test sessions would be a better recommendation.
In the construct validity analyses, the majority (five of the seven) of the FFT items successfully discriminated between the PA and SED groups. The only test items that did not prove discriminatory were the two flexibility measures. The scores on the FFT items are, for the most part, comparable to those from other studies that have used the same or similar tests. The majority of the evidence provides support for the construct validity of the FFT items.

The FFT battery is designed for use in a field setting; therefore, the practical utility of the battery is very important. Accordingly, the ability of the test battery to discriminate between physically active and sedentary groups is encouraging. The present study is cross-sectional in design, and in order to further validate the clinical utility of the FFT battery, longitudinal studies should be conducted.

References


