Development and Validation of a Functional Fitness Test for Community-Residing Older Adults

Roberta E. Rikli and C. Jessie Jones

Preventing or delaying the onset of physical frailty is an increasingly important goal because more individuals are living well into their 8th and 9th decades. We describe the development and validation of a functional fitness test battery that can assess the physiologic parameters that support physical mobility in older adults. The procedures involved in the test development were (a) developing a theoretical framework for the test items, (b) establishing an advisory panel of experts, (c) determining test selection criteria, (d) selecting the test items, and (e) establishing test reliability and validity. The complete battery consists of 6 items (and one alternative) designed to assess the physiologic parameters associated with independent functioning—lower and upper body strength, aerobic endurance, lower and upper body flexibility, and agility/dynamic balance. We also assessed body mass index as an estimate of body composition. We concluded that the tests met the established criteria for scientific rigor and feasibility for use in common community settings.

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

Finding ways to prevent or delay the onset of physical frailty in later years has become a major goal for gerontology researchers and practitioners throughout the world. According to previously published terminology, physical frailty is defined as the loss of physiologic reserve that increases the risk of disability (Buchner & Wagner, 1992). The proposed battery of tests has been developed to provide a means of assessing the key physiologic parameters that support functional mobility in older adults. Although physical decline during aging is due to multiple causes—a combination of biological aging, disease, and certain lifestyle patterns such as low levels of physical activity (American College of Sports Medicine [ACSM], 1998; Buchner & Wagner, 1992)—it is believed that much of this loss is preventable and even reversible through early detection of physical weakness and appropriate activity intervention (Gill, Williams, Richardson, & Tinetti, 1996; Guralnik, Ferrucci, Simonsick, Salive, & Wallace, 1995; Jackson et al., 1995; Lawrence & Jette, 1996; Morey, Pieper, & Cornoni-Huntley, 1998). Many independent older adults, often because of their sedentary lifestyles, are functioning dangerously close

Roberta E. Rikli and C. Jessie Jones are with the Division of Kinesiology and Health Promotion, California State University, Fullerton, Fullerton, CA 92834.
to their maximum capacity when performing normal everyday activities (e.g., climbing stairs, getting out of a chair, lifting objects; Chandler & Hadley, 1996; Evans, 1995b; Shephard, 1993; Shephard, 1997). Any further decline or small physical setback could easily cause them to move from independent to disabled status, in which assistance is needed for daily activities and the risk of falling and costly fall-related injuries is greatly increased.

Unfortunately, a limiting factor in evaluating and managing physical decline during aging has been the lack of suitable measurement tools to assess the underlying physical parameters associated with functional mobility—strength, endurance, flexibility, balance, and agility (Chandler & Hadley, 1996; Chodzko-Zajko, 1994; Fried et al., 1996; Rikli & Jones, 1997; Spirduso, 1995; Verbrugge & Jette, 1994). Most traditional protocols for assessing fitness (treadmill and cycle ergometer tests, bench step tests, 1RM strength tests, etc.) were developed and validated for younger people and generally are inappropriate or unsafe for the majority of older adults, particularly without medical release and close monitoring of medical conditions. Even the least demanding treadmill and aerobic step test protocols, for example, are too difficult for the majority of the community-residing older population, many of whom are quite sedentary and already have experienced substantial declines in physical capacity. In addition, many of these protocols require expensive equipment or extensive training for test technicians and are not, therefore, feasible for use in the clinical or community setting where fitness assessment of older adults generally takes place.

To date, most physical performance tools for assessing function in older adults were designed primarily to detect functional limitations at the behavioral level—that is, to assess difficulties in performing specific activities such as bathing and dressing oneself, walking, and stair climbing (Buchner, Guralnik, & Cress, 1995; Guralnik, Reuben, Buchner, & Ferrucci, 1995). The limitation of such instruments is that physical impairments often are not detected until late in the disability process—until they have progressed to the point at which they are manifested in overt loss of functional ability. Specifically lacking in the gerontology/exercise science literature are nonlaboratory tests that can detect the physiologic declines that are precursors to the loss of function, that is, declines that occur at the systems level, as in musculoskeletal, cardiovascular, and neurologic declines.

One of the few tests that have been developed specifically to assess these physiologic parameters (commonly referred to as fitness components) in older adults is the AAHPERD (American Alliance for Health, Physical Education, Recreation and Dance) functional fitness assessment battery (Osness et al., 1996). Although the AAHPERD test has been highly effective as a groundbreaking tool in the area of functional fitness assessment and in stimulating field-based research on older adults, it has a number of limitations. For example, the test currently does not include a measure of lower body muscle function, a fitness component that is especially critical in maintaining physical independence and preventing disability (Gill et al., 1996; Guralnik, Ferrucci, et al., 1995; Haskell & Phillips, 1995; Lawrence & Jette, 1996). In addition, the difficulty level on some of the AAHPERD test items limits their usability in some settings. For example, many older people cannot complete the 1/2-mile distance walk and therefore cannot be evaluated using this test. In fact, statistics indicate that 40% of community-residing adults over the age of 70 have difficulty walking even 1/4 mile (Select Committee on Aging, 1992).
Many older adults also cannot sit on the floor in a straight-leg position, as required for the AAHPERD flexibility test (Jones, Rikli, Max, & Guillermo, 1998). The performance and scoring limitations of the AAHPERD test (and others) make it impossible to achieve continuous-scale scoring across the wide continuum of ability levels typically found in the community-residing older adult population.

Another test developed specifically to assess mobility in older adults is the EPESE (Established Populations for Epidemiologic Studies of the Elderly) short battery of items to measure strength, balance, and gait speed (Guralnik et al., 1994). Although the tests in this battery have been successful in classifying large populations of community-residing older adults into broad categories of functional status, significant “ceiling” or “floor” effects on some of the items limit their ability to provide measurement data on a continuous scale across a wide range of ability levels. Nearly 50% of those tested, for example, reached perfect scores (i.e., 100% of those tested) on a tandem balance task (Guralnik et al., 1994; Seeman et al., 1994). Conversely, it was found that approximately 22% of the target population could not complete a 5-times chair-stand test of lower body strength (i.e., could not reach the baseline, or floor, requirements of the test). Although other performance tests have been developed to assess functional behavior in older adults (Buchner et al., 1995; Guralnik, Reuben, et al., 1995), we are aware of no other test batteries designed specifically to assess physiologic performance in the elderly, at least not in this country. Some significant work in this area is underway in other countries, however, such as Japan (Tanaka et al., 1995) and the Netherlands (Van Heuvelen, Kempen, Ormel, & Rispens, 1998).

**Goals and Procedures**

The purpose of this project was to develop and validate a battery of performance tests suitable for assessing the major underlying physical parameters associated with functional mobility in independent older adults, ages 60–90+. Of special concern was the development of reliable and valid tools that can measure performance on a continuous scale across a wide range of functional abilities—from the borderline frail to the highly fit. The procedures involved in developing the tests were (a) developing a theoretical framework for the test battery, (b) establishing an advisory panel of experts in the field, (c) determining test selection criteria, (d) selecting the test items, (e) establishing test reliability and validity, and (f) developing normative standards. The first five of these procedures will be discussed in this paper. The final stage of the process, that of developing normative scores, will be discussed in the following article in this journal.

**THEORETICAL FRAMEWORK**

An initial step in the test development was to define functional fitness within an ability/disability framework. Developing appropriate tools for assessing the physical status of older individuals requires an understanding of the progression leading to loss of function or, conversely, to its maintenance. Common models explaining the disabling process (Lawrence & Jette, 1996; Nagi, 1991) describe the usual progression as moving from physiologic impairment (decline in body systems—
muscular, cardiovascular, neurologic, etc.) to functional limitations (restrictions in physical behavior such as lifting, stooping, walking, or climbing stairs) to disability (the inability to perform normal daily activities such as bathing oneself, housework, or shopping).

Although traditional models (Nagi, 1991) indicate that all disability originates directly from disease or pathology (with disease leading to impairment, impairment to functional limitation, and functional limitation to disability—see Figure 1(a)), recent evidence suggests that disuse or physical inactivity can be just as responsible for the physical decline leading to disability and that perhaps the model should be revised as in Figure 1(b) (Chandler & Hadley, 1996; DiPietro, 1996; Morey et al., 1998). Such a modification has important implications both for prevention and for intervention strategies, as well as for developing relevant assessment tools. The growing body of evidence supporting the role of physical inactivity in the loss of function independent of the disease process comes from several large epidemiologic and other longitudinal studies (ACSM, 1997; Chandler & Hadley, 1996; Gill et al., 1996; Kaplan, Strawbridge, Camacho, & Cohen, 1993; Lacroix, Guralnik, Berkman, Wallace, & Satterfield, 1993; Lawrence & Jette, 1996; Mor et al., 1989; Morey et al., 1998; Seeman et al., 1995; Stewart et al., 1994).

The functional ability framework (Figure 2), based on the models shown in Figure 1, provided a useful guide during the process of defining functional fitness and in identifying relevant physical parameters for measurement. The framework depicts the progressive relationship among physical parameters, functional performance, and activity goals. The common activities in the far right column of Figure 2 (e.g., caring for personal needs, shopping, traveling) require the ability to perform the functions listed in column two (e.g., walking, stair climbing, lifting, reaching). These functions, in turn, require adequate reserve in the physical parameters

![Diagram](image)

**Figure 1.** (a) Nagi's (1991) model of the progression leading to disability and (b) an amended version suggesting that an inactive lifestyle can have comparable effects on the disabling process. Adapted from "Assessing Physical Performance in Independent Older Adults: Issues and Guidelines" by R.E. Rikli and C.J. Jones, 1997, *Journal of Aging and Physical Activity, 5*, p. 247. © 1997 by Human Kinetics. Adapted with permission.
identified in column one—muscular strength, endurance, flexibility, power, speed, agility, and balance—as well as an optimal (or at least manageable) body mass index. Supporting rationale for the physical fitness categories (column 1) relative to their importance in functional mobility has been well documented in major reports and publications (ACSM, 1997; ACSM, 1998; Bouchard, Shephard, & Stephens, 1994; Buchner, 1995; Chandler & Hadley, 1996; Hurley & Hagberg, 1998; Morey et al., 1998; U.S. Dept. of Health and Human Services, 1996).

Based on the framework outlined in Figure 2, functional fitness has been defined as having the physiologic capacity to perform normal everyday activities safely and independently without undue fatigue. In support of this definition, appropriate items on a functional fitness test battery would be those that reflect the physiologic attributes that support the behavioral functions necessary to perform activities of daily living. The phrase without undue fatigue is included in the definition to emphasize the importance of maintaining an adequate physiologic reserve. In fact, evidence suggests that a direct relationship between physiologic impairment (in strength, endurance, etc.) and functional limitation tends to exist only in the lower ranges of the performance spectrum, at least with respect to basic activities such as walking and stair climbing (Buchner & de Lateur, 1991; Buchner, Larson, Wagner, Koepsell, & de Lateur, 1996; Ferrucci et al., 1997; Jette, Assmann, Rooks, Harris, & Crawford, in press; Judge, Underwood, & Gennosa, 1993). Increasingly, however, gerontologists are recognizing the need for screening tools that can assess physical declines early enough not only to prevent disability—that is, to detect and treat declines before reserves have been depleted and the threshold

<table>
<thead>
<tr>
<th>PHYSICAL PARAMETERS</th>
<th>FUNCTIONS</th>
<th>ACTIVITY GOALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle strength/</td>
<td>Walking</td>
<td>Personal care</td>
</tr>
<tr>
<td>endurance</td>
<td>Stair climbing</td>
<td>Shopping/</td>
</tr>
<tr>
<td>Aerobic endurance</td>
<td>Standing up from chair</td>
<td>errands</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Lifting/reaching</td>
<td>Housework</td>
</tr>
<tr>
<td>Motor ability</td>
<td>Bending/kneeling</td>
<td>Gardening</td>
</tr>
<tr>
<td>power</td>
<td>Jogging/Running</td>
<td>Sports</td>
</tr>
<tr>
<td>speed/agility</td>
<td></td>
<td>Traveling</td>
</tr>
<tr>
<td>balance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body composition</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. A functional ability framework indicating the physiologic parameters associated with functions required for basic and advanced everyday activities. Adapted from “Assessing Physical Performance in Independent Older Adults: Issues and Guidelines” by R.E. Rikli and C.J. Jones, 1997, Journal of Aging and Physical Activity, 5, p. 246. © 1997 by Human Kinetics. Adapted with permission.
is reached at which ADL behavior is affected—but also to facilitate continued involvement in advanced exercise activities (Buchner & Wagner, 1992; Fried et al., 1996; Morey et al., 1998).

**ADVISORY PANEL OF EXPERTS**

Both local and national advisory panels were assembled to serve as consultants throughout the project. Local panel members, all exercise specialists in southern California, assisted with the development and pilot testing of the test items. The national advisory panel, comprising noted researchers and program leaders in the fields of gerontology or exercise science, reviewed the test development materials and the reliability and validity analysis procedures. All pilot testing for the project and other background research was conducted at the Ruby Gerontology Center at California State University, Fullerton, and in the surrounding communities.

**TEST SELECTION CRITERIA**

In choosing specific test items to assess the functional fitness components, consideration was given to two major goals: (a) the development of test protocols that meet acceptable scientific standards with respect to test reliability and validity and (b) the development of tests that would be easy to administer and feasible for use in common clinical and community settings in which the majority of assessment is most likely to take place. The following is a list of 12 criteria established for use as guidelines during the test development. It was agreed that all test items should

- Represent major functional fitness components—that is, key physiologic parameters associated with the functions required for independent living.
- Have acceptable test–retest reliability (≤.80).
- Have acceptable validity, with documentation to support at least two of the following: content validity, criterion validity, and construct (discriminant) validity. For acceptable criterion validity, correlations between the test item and the criterion measure were to be .70 or greater. For construct, or discriminant, validity, relevant group differences should be significant beyond the .01 level.
- Reflect normal age-related changes in physical performance.
- Be able to detect physical changes resulting from training or exercise.
- Be able to assess on a continuous scale across wide ranges of functional ability—from the low fit:borderline frail to the highly fit. The goal was to avoid "ceiling" and "floor" effects so that all, or most, participants would receive a score.
- Be easy to administer and score (by paraprofessionals and volunteer technicians, who often assist in administering tests).
- Require minimal equipment and space (can be administered in typical clinical and community settings).
- Be capable of being administered by oneself or a partner in the home setting.
- Be safe to perform without medical release for the majority of community-residing older adults.
- Be socially acceptable and meaningful.
- Be reasonably quick to administer, with individual testing time requiring no
more than 30–45 min. Groups of approximately 24 people should be able to complete the tests within 90 min, using five or six trained volunteer assistants.

**SELECTION OF TEST ITEMS**

The test selection process involved a thorough review of relevant literature, extensive input from advisory board members, and over 2 years of trial-and-error pilot testing to develop tests that would meet the established criteria. During the pilot testing, a major focus was on refining protocols to improve reliability, validity, safety, and feasibility with respect to time, equipment, and space requirements. Of special concern was determining the minimum number of trials needed to establish stable and reliable criterion measures. In addition, on several occasions, physicians and other health professionals observed participants as they were taking the tests for any signs of overexertion or other unsafe conditions. Ultimately, the test items listed in Table 1 were selected to assess the major components of functional fitness. The rationale for each test item is included as part of the validity discussion. Complete descriptions of test items are provided in Appendix A.

Although at first glance the test items may appear to be more behavioral in nature than physiologic, each item was selected because of its ability to reflect (in a reliable and valid way) one of the physical parameters of functional fitness. The

<table>
<thead>
<tr>
<th>Physical fitness parameters</th>
<th>Test items</th>
<th>Scoring protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower body strength</td>
<td>30-s chair stand</td>
<td>1 trial, with partial practice trial on same day</td>
</tr>
<tr>
<td>Upper body strength</td>
<td>arm curl</td>
<td>1 trial, with partial practice trial on same day</td>
</tr>
<tr>
<td>Lower body flexibility</td>
<td>chair sit-and-reach</td>
<td>best of 2 trials, following 2 practice trials on same day</td>
</tr>
<tr>
<td>Upper body flexibility</td>
<td>back scratch</td>
<td>best of 2 trials, following 2 practice trials on same day</td>
</tr>
<tr>
<td>Aerobic endurance</td>
<td>6-min walk</td>
<td>1 trial, following practice trial on a prior day</td>
</tr>
<tr>
<td>Alternative aerobic test</td>
<td>2-min step test</td>
<td>1 trial, following practice trial on a prior day</td>
</tr>
<tr>
<td>Motor agility/dynamic balance</td>
<td>8-ft up-and-go</td>
<td>best of 2 trials, following practice trial on same day</td>
</tr>
<tr>
<td>Body mass index</td>
<td>height and weight</td>
<td>1 measure of each</td>
</tr>
</tbody>
</table>

*For tests involving multiple trials, measurement theory suggests that the average score is normally more reliable than the best score. However, because the “best score” protocols for the three multiple-trial tests in this battery (back-scratch, chair sit-and-reach, and 8-ft up-and-go) were highly reliable (R > .90), we are recommending that the best score be used to save time in recording data.*
main purpose of the chair-stand test, for example, which was validated against 1RM (1-repetition maximum) leg-press strength, is to assess lower body strength, not just the ability to get out of a chair. Similarly, the 6-min walk is included to assess aerobic endurance, not simply walking ability.

ESTABLISHING TEST–RETEST RELIABILITY

Participants. Eighty-two older adults (48 women and 34 men; mean age = 71.8 years, SD = 6.9) were solicited from a nearby senior housing complex and from enrollees in a university-sponsored exercise program to participate in testing to establish the reliability of the selected items. Each participant signed an informed consent and completed a written questionnaire requesting information about his or her health, functional status, and physical activity level. The criteria for inclusion in the study were that participants be over the age of 60, be community residing, be ambulatory without the use of assistive devices, have no medical conditions that would be contraindicated for submaximal testing (ACSM, 1995), and not have been advised by their physicians to refrain from exercise. Participant characteristics for the reliability studies are presented in Table 2.

Reliability Procedures. All of the tests (except for the 6-min walk) were administered on the same day, with retesting occurring 2 to 5 days later. The 6-min walk test was administered separately approximately 3 weeks later, again with 2 to 5 days separating the test trials. Prior to all testing, participants performed 8 min of warm-up and stretching exercises. All tests were administered by trained graduate students and volunteer senior technicians. During the training, technicians practiced on each other until they demonstrated proper procedures to the project coordinators. So that test objectivity would be reflected in the reliability estimates, different technicians were used to conduct the tests on each of the two test days. Day 2 technicians were not aware of the scores obtained on Day 1. Because data indicate that significant increases can occur in mean scores from Day 1 to Day 2 on two of the test items—the 6-min walk (Rikli & Jones, 1998) and the 2-min step-in-place test (Dugas, 1996)—participants received three trials on both of these tests 2 to 5

| Table 2 Descriptive Characteristics of Study Participants—Means and Standard Deviations |
|-----------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                   | Men (n = 34)    | Women (n = 48)  |
|                                   | M               | SD              | M               | SD              |
| Age (years)                       | 72.6            | 6.6             | 69.1            | 5.1             |
| Height (cm)                       | 177.0           | 7.4             | 163.1           | 5.8             |
| Weight (kg)                       | 83.1            | 16.6            | 71.2            | 14.3            |
| Number of chronic conditions      | 1.9             | 1.8             | 1.8             | 1.6             |
| Number of medicationsa            | 2.0             | 1.7             | 1.9             | 2.1             |
| Number of years of education      | 18.6            | 2.1             | 16.0            | 1.9             |

aDifferent prescription medications taken each day.
days apart, with the first trial treated as a practice trial. Although a significant trial, or “practice,” effect (increase in scores from one day to the next) does not always have a large impact on the relative reliability of test–retest scores (i.e., the individual positioning of scores relative to other scores in the group), it does affect absolute scoring consistency and the stability reliability of the test. Scoring consistency (the lack of a practice effect) is of particular concern in intervention studies, in which stable baseline measures are required for accurate interpretation of treatment effects.

**Data Analysis.** Test–retest reliability for all test items was established by calculating the intraclass correlation coefficient ($R$), using a one-way ANOVA model appropriate for estimating what the reliability would be for a single test (Baumgartner & Jackson, 1995). The one-way ANOVA also provides information on the amount of change in scores, if any, from one day to the next. A lack of significant change indicates scoring stability across trials.

**Results.** Intraclass reliability values ($R$) and 95% confidence intervals (CI) for all tests are presented in Table 3. The $R$ values for the test items ranged from .80 to .98—with a majority of the values being .90 or above, indicating that the tests have good relative reliability across trials. The $R$ values in Table 3 reflect intraclass correlations between Trials 1 and 2 on all items except the 6-min walk and 2-min step test. On these tests, for which Trial 1 was treated as a practice trial, the $R$ values reflect intraclass correlations between Trials 2 and 3. In an earlier study on step-test performance in older adults, Dugas (1996) found intraclass correlations of around .50 between Trials 1 and 2 but correlations ranging from .85 to .97 between Trials 2 and 3. The same data also revealed a significant increase in mean scores from Day 1 to Day 2 ($p < .001$), but not between Trials (Days) 2 and 3, thus suggesting that treating Trial 1 as a practice trial and Trial 2 as the official test trial results in superior relative reliability, as well as stability reliability.

Similarly, as reported elsewhere (Rikli & Jones, 1998), a significant increase was found in 6-min walk performance from Day 1 to Day 2, but not between Days 2 and 3. However, on the 6-min walk, the intraclass correlations (relative reliability) of .88 to .91 between Trials 1 and 2 were nearly as high as the .91 to .97 correlations

<table>
<thead>
<tr>
<th>Test item</th>
<th>All Participants</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$ (CI)</td>
<td>$R$ (CI)</td>
<td>$R$ (CI)</td>
</tr>
<tr>
<td>30-s chair stand</td>
<td>.89 (.79–.93)</td>
<td>.86 (.77–.90)</td>
<td>.92 (.87–.95)</td>
</tr>
<tr>
<td>Arm curl</td>
<td>.81 (.72–.88)</td>
<td>.81 (.66–.90)</td>
<td>.80 (.67–.89)</td>
</tr>
<tr>
<td>6-min walk</td>
<td>.94 (.90–.96)</td>
<td>.97 (.92–.99)</td>
<td>.91 (.84–.95)</td>
</tr>
<tr>
<td>2-min step test</td>
<td>.90 (.84–.93)</td>
<td>.90 (.80–.95)</td>
<td>.89 (.83–.94)</td>
</tr>
<tr>
<td>Chair sit-and-reach</td>
<td>.95 (.92–.97)</td>
<td>.92 (.85–.96)</td>
<td>.96 (.93–.98)</td>
</tr>
<tr>
<td>Back scratch</td>
<td>.96 (.94–.98)</td>
<td>.96 (.93–.98)</td>
<td>.92 (.86–.96)</td>
</tr>
<tr>
<td>8-ft up-and-go</td>
<td>.95 (.92–.97)</td>
<td>.98 (.96–.99)</td>
<td>.90 (.83–.95)</td>
</tr>
</tbody>
</table>
between Trials 2 and 3. On all other test items, ANOVA results revealed no significant changes in scores from Day 1 testing to Day 2, thus indicating that the tests also have absolute, or stability, reliability across trials.

ESTABLISHING TEST VALIDITY

Content validity (or logical validity, as it usually is called when referring to performance measures) is the degree to which a test (or test battery) reflects a defined universe or domain of content (American Psychological Association, 1985). For performance measures, content validity is typically established by examining, through literature review or through expert opinion, the degree to which a test measures the capacity (or capacities) that it is intended to measure. The content relevance (or logical validity) of each item in the battery of tests is discussed in the next section.

Criterion validity represents the degree to which a test correlates with a criterion measure that is already known to be valid. The criterion-related validity of the test items was estimated by calculating an interclass correlation coefficient (Pearson’s r) between the test item scores and scores on a criterion measure of performance, whenever a suitable criterion could be identified. Validation procedures are discussed separately for each test item and are summarized in Table 4.

Construct, or discriminant, validity represents the degree to which a test measures a particular construct (an attribute that exists in theory but cannot be directly measured) such as “functional fitness.” Assessing construct validity typically involves comparing test results of two or more groups that have known or presumed differences with respect to the construct of interest (Morrow, Jackson, Disch, & Mood, 1995). A test’s ability to discriminate among such groups is an indication of its construct, or discriminant, validity. To assess the construct validity of the tests in this battery, comparisons were made between older adults who were regular exercisers and those who were not regularly active, with the assumption that exercisers would possess higher levels of fitness than nonexercisers. Comparisons were also made across different age groups of individuals (those in their 60s, 70s, and 80s) in which performance is generally expected to decline.

Participants in the construct validity studies (except for the 6-min walk) consisted of 190 male and female residents of a nearby retirement housing complex (mean age = 76.2, SD = 6.7). Participants in the 6-min walk study were 77 volunteer men and women (mean age = 73.1, SD = 7.2) solicited from another retirement housing complex and from a university-sponsored exercise program. Each participant signed an informed consent and completed a written questionnaire requesting information about his or her age, height, weight, health status, and physical activity level. Participants were ambulatory without the use of assistive devices and not suffering from joint pain, unstable cardiovascular disease, or other medical condition that would be contraindicated for submaximal testing (ACSM, 1995).

Test Items

Following is a general description of each test item, along with supporting evidence of the item’s validity as an indicator of functional fitness. Full descriptions of each test item and instructions for administering the tests are presented in Appendix A.
Table 4  Criterion Validity Estimates—Correlations ($r$) Between Test Items and Criterion Measures

<table>
<thead>
<tr>
<th>Test item</th>
<th>Criterion measure</th>
<th>All Participants</th>
<th></th>
<th>Men</th>
<th></th>
<th>Women</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-s chair stand</td>
<td>1RM leg press</td>
<td>.77  89</td>
<td>.78  40</td>
<td>.71  49</td>
<td></td>
<td></td>
<td>Jones &amp; Rikli, in press</td>
</tr>
<tr>
<td>Arm curl</td>
<td>combined 1RM chest press, biceps, and upper back$^a$</td>
<td>.81  20</td>
<td>.78  24</td>
<td></td>
<td></td>
<td></td>
<td>James et al., 1998</td>
</tr>
<tr>
<td>6-min walk</td>
<td>time on treadmill to 85% max heart rate</td>
<td>.78  37</td>
<td>.82  17</td>
<td>.71  20</td>
<td></td>
<td></td>
<td>Rikli &amp; Jones, 1998</td>
</tr>
<tr>
<td>2-min step</td>
<td>1-mile walk time</td>
<td>.73  24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dugas, 1996</td>
</tr>
<tr>
<td></td>
<td>time on treadmill to 85% max heart rate</td>
<td>.74  25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Johnston, 1998</td>
</tr>
<tr>
<td>Chair sit-and-reach</td>
<td>goniometer-measured hamstring flexibility</td>
<td>.83  80</td>
<td>.76  32</td>
<td>.86  48</td>
<td></td>
<td></td>
<td>Jones et al., 1998</td>
</tr>
<tr>
<td>Back scratch</td>
<td>no single criterion available</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>considered to be best measure of overall shoulder flexibility</td>
</tr>
<tr>
<td>8-ft up-and-go</td>
<td>no single criterion available</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>considered to be good measure of combined physiologic attributes (power, speed, agility, and balance)</td>
</tr>
</tbody>
</table>

$^a$Because of the difference in test protocols for men and women (5-lb vs. 8-lb weights), it was inappropriate to combine them for the correlation analysis.
Because the tests were designed primarily for clinical and community settings in the United States, nonmetric units (traditional U.S. measurements) have been used to describe test protocols and scoring procedures. Guidelines for converting the nonmetric units to the metric system are included in Appendix B.

30-SECOND CHAIR STAND

**Description.** The 30-s chair-stand test, selected to reflect lower body strength, involves counting the number of times within 30 s that an individual can rise to a full stand from a seated position, without pushing off with the arms (see illustration in Appendix A). The test is a modified version of the timed-stands test published by Csuka and McCarty (1985), which involved recording the amount of time required to stand 10 times. The modification was made in order to avoid the floor effect associated with the original version, thus increasing the evaluation range and discrimination ability of the test. The original test protocol requiring 10 stands makes it impossible for a sizable portion of older adults to complete the test. As indicated earlier, nearly 22% of the 5,000 community residents over the age of 71 in the EPESE studies could not complete a test protocol involving even five chair stands (Guralnik et al., 1994). Changing the protocol to a prescribed number of seconds rather than a prescribed number of stands makes it possible for all individuals to receive a score, even though the score may be zero in extreme cases.

**Justification/Content Relevance.** Lower body muscular integrity has been well established as a major factor in maintaining functional mobility and preventing or delaying the onset of disability (Gill et al., 1996; Guralnik, Ferrucci, et al., 1995; Haskell & Phillips, 1995; Lawrence & Jette, 1996; Pendergast, Fisher, & Calkins, 1993). Declines in lower body strength have been associated with the deterioration of a number of performance variables such as gait, stair climbing, rising from a chair, and balance (Bassey et al., 1992; Bohannon, 1995; Brown, Sinacore, & Host, 1995; Evans, 1995a; Fialarone et al., 1990; Judge, 1993). Past studies have shown that chair-stand performance, a common method of assessing lower body strength in older adults, correlates reasonably well with major criterion indicators of lower body strength (Cybex II knee extensor and knee flexor strength and other functional measures), stair-climbing ability, walking speed, and risk of falling (Bohannon, 1995; Csuka & McCarty, 1985). In addition, chair-stand performance has been found to detect normal age-related declines (Csuka & McCarty) and discriminate between fallers and nonfallers (MacRae, Lacourse, & Moldavon, 1992) and is associated with risk of falling (Alexander, Schultz, & Warwick, 1991; Tinetti, Speechley, & Ginter, 1988). Furthermore, the chair stand has been found to be safe (Guralnik et al., 1994) and capable of detecting the effects of physical training in older adults (McMurdo & Rennie, 1993).

**Additional Test Validation.** To test the criterion validity of the modified chair-stand test, the scores of 76 older adults (42 women, 34 men; mean age = 70.5, SD = 5.5) were compared with 1RM leg-press strength (Jones & Rikli, in press). The leg press, a multiple-joint exercise involving hip extension, knee extension, and ankle plantar flexion, is considered an especially good criterion measure of lower body strength in older adults because it reflects many common daily activities such as rising from a chair or getting out of a tub or a car. The moderately high correlation (see Table 4) between chair-stand performance and maximum leg-press performance for both men and women (r = .78 and .71, respectively) supports the criterion
Table 5  Age-Group Comparisons for 60-, 70-, and 80-Year-Olds on All Test Items

<table>
<thead>
<tr>
<th>Test item</th>
<th>60–69 yr (n = 32, 91% female)</th>
<th>70–79 yr (n = 96, 84% female)</th>
<th>80–89 yr (n = 62, 81% female)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-s chair stand (no. in 30 s)</td>
<td>14.0 (2.4)</td>
<td>12.9 (3.0)</td>
<td>11.9 (3.6)</td>
<td>.013*</td>
</tr>
<tr>
<td>Arm curl (no. in 30 s)</td>
<td>19.8 (4.1)</td>
<td>18.2 (3.9)</td>
<td>16.5 (4.1)</td>
<td>.007*</td>
</tr>
<tr>
<td>6-min walk (total yd)</td>
<td>677.8 (95.0)</td>
<td>621.0 (82.4)</td>
<td>550.1 (86.7)</td>
<td>.0001*</td>
</tr>
<tr>
<td>2-min step (no. in 2 min)</td>
<td>100.4 (9.0)</td>
<td>92.6 (16.0)</td>
<td>83.5 (22.6)</td>
<td>.0001*</td>
</tr>
<tr>
<td>Chair sit-and-reach (in. from toe)</td>
<td>-0.4 (5.4)</td>
<td>-0.4 (6.2)</td>
<td>-3.3 (6.0)</td>
<td>.018*</td>
</tr>
<tr>
<td>Back scratch (in. from middle fingers)</td>
<td>1.0 (2.0)</td>
<td>-0.4 (3.0)</td>
<td>-1.8 (4.4)</td>
<td>.0009*</td>
</tr>
<tr>
<td>8-ft up-and-go (s)</td>
<td>5.2 (0.6)</td>
<td>6.1 (1.2)</td>
<td>7.1 (2.0)</td>
<td>.0001*</td>
</tr>
</tbody>
</table>

Note. Group means, (standard deviations), and p values from one-way ANOVAs are presented. Mean ages and standard deviations for the 60-, 70-, and 80-year-old groups were 65.9 (3.5), 75.1 (2.9), and 83.3 (3.0), respectively.

*Post hoc analysis revealed that all three age groups differed significantly. Only the 80-year-olds were significantly different from the other age groups.

validity of the test. In support of the construct, or discriminant, validity of the test, chair-stand scores were found to decline across age (see Table 5) and were lower for low-active participants compared with high-active participants (see Table 6).

ARM CURL

**Description.** The arm-curl test, a measure of upper body strength, involves determining the number of times a hand weight (5 lb for women, 8 lb for men) can be curled through a full range of motion in 30 s. This test item is similar to the AAHPERD arm-curl test (Osness et al., 1996), with two exceptions: a change in the prescribed weight for females (from a 4-lb to a 5-lb weight) and a change in arm position during the curl-up phase of the movement. The weight for females was changed because a market survey revealed that 5-lb weights were much easier to obtain than 4-lb weights (which contributes to test feasibility in community programs) and, more important, because upper body strength in women tends to be about 60% of that in men (Sperling, 1980), so the 5-lb weights would improve the representativeness of the test protocol across genders. The prescribed arm-position
Table 6  Comparisons for High-Active vs. Low-Active Participants

<table>
<thead>
<tr>
<th>Test item</th>
<th>High-active (n = 136)</th>
<th>Low-active (n = 47)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-s chair stand (no. in 30 s)</td>
<td>13.3 (2.8)</td>
<td>10.8 (3.6)</td>
<td>.0001</td>
</tr>
<tr>
<td>Arm curl (no. in 30 s)</td>
<td>18.7 (4.0)</td>
<td>15.5 (3.7)</td>
<td>.0001</td>
</tr>
<tr>
<td>6-min walk(^a) (total yd)</td>
<td>647.6 (81.5)</td>
<td>513.2 (77.9)</td>
<td>.0001</td>
</tr>
<tr>
<td>2-min step (no. in 2 min)</td>
<td>95.8 (15.7)</td>
<td>72.8 (18.4)</td>
<td>.0001</td>
</tr>
<tr>
<td>Chair sit-and-reach (in. from toe)</td>
<td>-0.6 (6.0)</td>
<td>-3.8 (6.6)</td>
<td>.007</td>
</tr>
<tr>
<td>Back scratch (in. from middle fingers)</td>
<td>-0.3 (3.4)</td>
<td>-2.1 (3.8)</td>
<td>.005</td>
</tr>
<tr>
<td>8-ft up-and-go (s)</td>
<td>6.0 (1.3)</td>
<td>7.1 (2.1)</td>
<td>.0001</td>
</tr>
</tbody>
</table>

*Note. High-active participants were those who, according to self-report, were regular participants (at least three times per week) for at least 1 year in physical exercise or activity that was strenuous enough to cause a noticeable increase in heart rate, breathing, or perspiration. Low-active participants were new enrollees in exercise classes who had not participated in regular exercise for at least 5 years. Mean ages: high-active = 75.6 yr (SD = 6.6); low-active = 77.7 yr (SD = 6.9).

\(^a\)Group size for the 6-min walk was less than for the other measures, with 59 participants in the high-active and 17 in the low-active groups. Because of the additional time and energy demands of the 6-min walk studies, the research on this test was conducted on different days with different subject populations. See Rikli and Jones (1998) for more details.

The protocol includes holding the weight in a handshake grip at full extension (to the side of the chair), then supinating during flexion so that the palm of the hand faces the biceps at full flexion. The rationale for changing from the AAHPERD protocol—that of keeping the handshake grip throughout the entire curl—to a palm-up protocol during flexion was to better engage the muscles of the upper arm and to enhance the effectiveness of the biceps tendon relative to muscle action. Results of pilot studies indicated that participants with arthritis were less bothered performing the arm-curl test described here than the maximum-grip strength protocol commonly used as a strength measure in other studies.

**Justification/Content Relevance.** Upper body function, including arm strength and endurance, is important in executing many normal everyday activities such as household chores, carrying groceries, lifting a suitcase, and picking up grandchildren. Although impairments in lower body function have been found to be stronger
predictors of the initial onset of disability—that is, limitations in activities required for functioning independently in the community—both lower body and upper body impairments are associated with dysfunction in later years, particularly ADL (activities of daily living) dysfunction—the inability to perform personal care activities (Lawrence & Jette, 1996). Although upper body strength generally declines with age (Frontera, Hughes, Lutz, & Evans, 1991), it also has been found to be modifiable through exercise (McCartney, Mc Kelvie, Martin, Sale, & MacDougall, 1993; Moritani & de Vries, 1980).

**Additional Validation of Test Item.** Data collected at Fullerton (James, Rikli, & Jones, 1998) indicate that scores on the 30-s arm-curl test had only a moderate correlation (.62 for men, .68 for women) with 1RM biceps strength, which was assessed on a Keiser pneumatic biceps machine. Of greater importance, however, relative to overall functional fitness, is the finding that 30-s arm-curl performance is a moderate to good indicator of overall upper body strength. Correlations between arm-curl performance and overall upper body strength, as indicated by combined 1RM biceps, 1RM chest-press, and 1RM upper back (seated row) strength measures (all taken on Keiser strength equipment), were found to be .81 for men and .78 for women (see Table 4). These findings suggest that the 30-s arm curl is at least moderately effective as a predictor of both biceps strength and overall upper body strength. Also, as indicated in Tables 5 and 6, the 30-s arm curl is able to significantly discriminate across different age groups (60s, 70s, and 80s) and between high-active and low-active individuals.

**6-MINUTE WALK**

**Description.** The 6-min walk test, a measure of aerobic endurance, involves determining the maximum distance (yd) that can be walked in 6 min along a 50-yd rectangular course (see diagram in Appendix A). The rationale for standardizing the time (6 min) instead of the distance (such as 1 mile or 1/2 mile—traditionally the more common protocol in run/walk tests) is to increase the assessment range and reduce the floor effects (baseline difficulty level) of the test. On a timed test (such as the 6-min test), scores can be obtained for the full range of ability levels—for the less fit individuals who can walk only a few yards, as well as for the highly fit person who can cover several hundred yards within the time allotted. Because reports have shown that considerable numbers of community-dwelling older adults have difficulty walking even 1/4 mile (Select Committee on Aging, 1992), tests with prescribed distances could be prohibitive to many of the elderly.

**Justification/Content Relevance.** An adequate level of aerobic endurance during aging is necessary in order to perform many everyday activities such as walking, shopping, or performing recreational or sports activities (Shephard, 1993). Although it has been estimated that a VO\(_{2}\)\(_{\text{max}}\) (physical work capacity) of 15–16 ml \cdot kg\(^{-1}\) \cdot min\(^{-1}\) is necessary to maintain independent living status, declines associated with inactive lifestyles often progress below this point before age 80 (Blair, Brill, & Kohl, 1989). Considerably higher levels of aerobic capacity, of course, are needed for higher level functioning—sports, hiking, traveling, and so on. A number of studies have indicated that distance walk/run tests of various dimensions (1 mile, 12 min, and 1/2 mile) are reasonably good indicators of aerobic endurance in both young adults (Cooper, 1968; Disch, Frankiewicz, & Jackson, 1975; Kline et al., 1987) and high-functioning older adults (Bravo et al., 1994;
Fenstermaker, Plowman, & Looney, 1992; Warren, Dotson, Nieman, & Butterworth, 1993). Shorter walking tests, such as a 5- or 6-min walk, have been used successfully to evaluate physical endurance in patients with various medical conditions (Bittner et al., 1993; Guyatt et al., 1985; Peloquin, Gauthier, Bravo, Lacombe, & Billiard, 1998) but had not, prior to this project, been validated on a generally healthy, community-residing population of older adults (Rikli & Jones, 1998).

**Additional Validation.** The concurrent validity of the 6-min walk was estimated by comparing 6-min walk scores with treadmill performance using a modified Balke-graded exercise test, a common criterion measure of physical endurance (ACSM, 1991). Physical endurance is defined as the ability to maintain submaximal exercise for an extended period of time (ACSM, 1997). The validation study, described in detail in an earlier publication (Rikli & Jones, 1998), involved correlating 6-min walk performance with time on treadmill to 85% predicted maximum heart rate for 76 men and women (mean age = 73.1, SD = 7.2) solicited from a nearby senior housing project and from enrollees in a university-sponsored exercise program. As indicated in Table 4, the correlations between the 6-min walk and the criterion measure ranged from .71 to .82, indicating moderately good support for the criterion validity of the test. In addition, the test’s ability to discriminate across age levels (60s, 70s, and 80s) and activity levels (high-active vs. low-active) lends support to the construct validity of the test (see Tables 5 and 6).

**2-MINUTE STEP (ALTERNATIVE AEROBIC TEST)**

**Description.** A 2-min step test is included in the test battery as an alternative aerobic endurance test, to be used when space prohibits conducting the 6-min walk. The 2-min step-in-place test might be considered a “self-paced” version of numerous other step tests, all of which require a specified stepping cadence. Pilot testing revealed that many older adults cannot or will not maintain a specified cadence. The 2-min step-in-place protocol involves determining the number of times, within 2 min, that an individual can step in place, raising the knees to a height halfway between the iliac crest (hip bone) and the middle of the patella (knee cap; see diagram in Appendix A). Data collected as part of a master’s thesis at California State University, Fullerton (Johnston, 1998), indicated that the average rate of perceived exertion (RPE) reported during the step test was fairly comparable to that reported during the 6-min walk test. On a scale of 1 to 20, the average RPE reported during the step test was 13.9, versus 13.6 during the 6-min walk.

**Justification/Validation.** The 2-min step-in-place test has been found to correlate reasonably well with other common measures of aerobic endurance. Dugas (1996) found a moderate correlation (r = .73) between 2-min step-test scores and 1-mile walk times in 24 older men and women (mean age = 69.6, SD = 6.5). In another study (Johnston, 1998), a similar correlation (r = .74) was found between 2-min step-test scores and treadmill performance in 25 men and women (mean age = 72.1, SD = 6.2). Treadmill performance represented time on treadmill to 85% predicted maximum heart rate, using a modified Balke-graded exercise test (ACSM, 1991). As indicated in Tables 5 and 6, the 2-min step test also revealed significant differences between age groups (60s, 70s, and 80s) and between exercisers and nonexercisers.
CHAIR SIT-AND-REACH

**Description.** The chair sit-and-reach, a measure of lower body (primarily hamstring) flexibility, is a modified version of other “floor” sit-and-reach tests that have appeared in a number of test batteries including the YMCA battery (Golding, Myers, & Sinning, 1989), the Fitnessgram (Morrow, Falls, & Kohl, 1994), and AAHPERD’s functional fitness test battery (Osness et al., 1996). In the revised protocol (Jones et al., 1998), the participant sits on the front edge of a chair and extends one leg straight out in front of the hip, with foot flexed and heel resting on the floor (the other leg is bent, foot flat on the floor). The object, the same as in the floor sit-and-reach, is to reach as far forward as possible toward (or past) the toes (see illustration in Appendix A).

The rationale for developing the chair protocol was to increase the rate of participation in older people and to decrease the risk of injury. Due to various conditions (e.g., obesity, lower back pain, lower body weakness, hip and knee replacements, severely reduced flexibility), many older people find it difficult or impossible to get down to and up from a floor position. We have also found that some older adults, possibly because of weak abdominal muscles, as well as tight hamstrings, cannot hold a sitting position on a flat surface (especially with both legs extended) and will start to fall backward during testing. The rationale for keeping one leg straight and one bent during the testing, as opposed to both legs straight, was based on evidence that the simultaneous stretching of both hamstrings increases the risk of back injury because of the excessive posterior disc compression that occurs in this position (Cailliet, 1988).

**Justification/Content Relevance.** Insufficient range of motion in the hip and insufficient hamstring flexibility are associated with low-back pain, increased susceptibility to musculoskeletal injury, gait limitation, and increased risk of falling in older adults (ACSM, 1995; Grabiner, Koh, Lundin, & Jahnigen, 1993; Kendall, McCreary, & Provance, 1993; Liemohn, Snodgrass, & Sharpe, 1988). Because of the importance of hamstring flexibility, some version of a sit-and-reach test has been included in nearly every fitness battery published in recent years.

**Additional Validation.** In order to assess the criterion validity of the chair sit-and-reach, scores were compared with a common “gold standard” criterion—that of goniometer-measured hamstring flexibility (American Academy of Orthopaedic Surgeons, 1966)—as well as other established measures of hamstring flexibility: a two-leg floor sit-and-reach (SR) and a one-leg back-saver sit-and-reach (BSR). As described in Jones et al. (1998), the correlation between chair sit-and-reach scores and the goniometer-measured criterion was .81 for women (n = 48) and .76 for men (n = 32) (mean age = 74.2, SD = 6.1). Interestingly, even for the subset of older adults in this study who were able to get into a proper floor sit-and-reach position to be assessed, the chair protocol for assessing hamstring flexibility was slightly better correlated with goniometer-measured hamstring flexibility than was either the two-leg or the one-leg floor sit-and-reach protocol. As indicated in Table 5, scores from the chair sit-and-reach test discriminated significantly between 70- and 80-year-olds but not between 60- and 70-year-olds. As can be seen in Table 6, chair sit-and-reach scores were significantly better for exercising individuals than for nonexercisers.
BACK SCRATCH

Description. The back-scratch test is a modified version of the Apley’s scratch test, which has been used for years by therapists and orthopedic physicians as a quick way of evaluating overall shoulder range of motion (Gross, Fetto, & Rosen, 1996; Hoppenfeld, 1976; Magee, 1992; Starkey & Ryan, 1996). The test, which involves a combination of shoulder abduction, adduction, and internal and external rotation, involves measuring the distance between (or the overlap of) the middle fingers behind the back. The Apley’s protocol, which involves reaching behind the head with one hand and behind the back with the other toward a specified anatomical point on the opposite scapula, was revised slightly to involve simply trying to touch the middle fingers of both hands together behind the back (see illustration in Appendix A). The assessment protocol was changed to provide a simpler and more quantifiable method of measuring shoulder range of motion in the field setting.

Justification/Content Relevance. Adequate shoulder flexibility is required for a number of everyday functions such as reaching behind the head (external rotation and abduction) to comb one’s hair, zip a back zipper, or put on over-the-head garments (T-shirts, sweatshirts, etc.). Reaching behind the lower back (internal rotation and adduction) is required for common functions such as removing a wallet from a back pocket, fastening a bra, or washing one’s back. Reduced range of motion in the shoulder girdle can result in pain and postural instability (Magee, 1992) and has been found to cause significant disability in as much as 30% of the healthy adult population over the age of 65 (Chakravarty & Webley, 1993).

Additional Validation. Although there is no single criterion measure for this test, it is considered by experts in the field to be a valid assessment of overall shoulder range of motion (Gross et al., 1996; Hoppenfeld, 1976; Starkey & Ryan, 1996). The hand-over-the-shoulder-and-behind-the-head motion assesses shoulder flexion, abduction, and external rotation; the hand-behind-the-back position involves shoulder extension, adduction, and internal rotation. As seen in Tables 5 and 6, the back-scratch test can significantly discriminate between age groups (60s, 70s, and 80s) and between activity levels (exercisers vs. nonexercisers).

8-FOOT UP-AND-GO

Description. The 8-ft up-and-go test, a composite measure involving power, speed, agility, and dynamic balance (an integration of neuromuscular system parameters), is a modified version of a previously published 3-m, timed up-and-go protocol (Podsiadlo & Richardson, 1991). The 8-ft up-and-go involves getting out of a chair, walking 8 ft to and around a cone (or other marker), and returning to the chair in the shortest time possible. The main purpose in changing the distance from 3 m (9.84 ft) to 8 ft was to increase the feasibility of administering this test in areas with limited space, particularly in the home setting. Pilot studies indicated that it was generally possible in most homes to conduct an 8-ft test (including room for turning) but often not possible to find the space needed for the 3-m distance. Data also indicated that very little scoring accuracy was lost in shortening the test from 3 m to 8 feet. Overall test–retest reliability for the 8-ft up-and-go was .95 (see Table 3), versus a reported .99 for the 3-m test (Podsiadlo & Richardson, 1991).
Justification/Content Relevance. The timed up-and-go test, which involves getting up from a seated position and moving quickly to a target location, reflects common mobility and gait maneuvers required in independent living, such as getting up from one’s seat to get off a bus in a timely and safe manner or getting up quickly to answer the phone or the door, to go to the bathroom, or to tend to something in the kitchen. Although there is no one “gold standard” criterion measure for this test, it has been found to be significantly related to the Berg balance scale ($r = -.81$), to gait speed ($r = .61$), and to the Barthel index of ADLs ($r = .78$)—a composite measure involving transfer actions (e.g., getting in and out of a tub), walking, and stair climbing (Podsiadlo & Richardson, 1991). Up-and-go performance has also been found to be a good predictor of recurrent falling, can discriminate among various functional categories, and is responsive to changes resulting from activity interventions (Podsiadlo & Richardson; Tinetti, Williams, & Mayewski, 1986).

Additional Validation. Studies conducted at Fullerton indicate that 8-ft up-and-go performance is an excellent discriminator among various age groups, with performance declining significantly from the 60s to the 70s to the 80s (Table 5). The test also resulted in significantly better scores for individuals who were regular exercisers than for those who were not (Table 6).

BODY MASS INDEX

Although body mass index (BMI), a weight/height ratio that correlates with body fat, is not a measure that was validated in this study, we are recommending that it be included as part of the functional fitness test battery because of previous evidence showing its role in maintaining functional mobility. Studies clearly show that individuals with high BMIs (or in some cases very low BMIs) are more likely to be disabled in later years than are people with normal body mass ratings (Galanos, Peiper, Comoni-Huntley, Bales, & Fillenbaum, 1994; Harris, Kovar, Suzman, Kleinman, & Feldman, 1989; Losonczy et al., 1995). High BMIs also are associated with numerous health problems including hypertension, heart disease, and Type II diabetes (U.S. Dept. of Health and Human Services, 1996).

BMI is determined by dividing weight in kilograms by height in meters squared: $\text{BMI} = \frac{\text{kg}}{\text{m}^2}$. An alternative formula involves multiplying weight in pounds by 703 and dividing by height in inches squared: $\text{BMI} = \frac{(\text{lb} + 703)}{\text{in.}^2}$. An even easier way to calculate BMI is to use a conversion table such as the one that is available on the Website of Shape Up America (http://www.shapeup.org), a national initiative promoting healthy body weight. Although it is not clear what the “ideal” BMI in older adults is, those with BMI ratings between 20 and 25 are generally considered to be in the normal range, whereas those with BMI ratings above 25 are considered overweight and at increased risk for disease and loss of functional mobility (Shephard, 1997). BMIs below 20, especially in older adults, can be an indication of loss of muscle mass and bone tissue.

Summary

Based on the studies conducted as part of this project, on feedback from the expert reviewers, and on the comments received from hundreds of people throughout the
nation who have been participating in these tests as part of the data collection to establish national norms, we believe that this battery of functional fitness tests has been successful in meeting the established criteria. Evidence has been presented to support acceptable test-retest reliability (.80 < R < .98) and at least two (sometimes three) types of validity for each test. The content (logical) validity of each test was demonstrated through literature review and expert opinion. As an indication of construct validity, all tests were able to significantly discriminate between individuals who were regular exercisers and those who were not, with exercisers consistently demonstrating superior performance, and, with only one exception, the tests were successful in reflecting expected performance declines across each decade of age, from the 60s to the 70s to the 80s. On the chair sit-and-reach, significant performance declines were detected between the 70- and 80-year-olds but not between those in their 60s and 70s. In addition, empirical evidence was presented to support the criterion validity of five of the test items (chair stand, arm curl, 6-min walk, 2-min step, and chair sit-and-reach). Expert opinion or a composite of former study results was presented to support the validity of the back-scratch and 8-ft up-and-go tests, tests for which no identifiable criterion standards exist.

The tests have been found to be successful in measuring performance on a continuous scale across a wide range of ability levels, without significant floor or ceiling effects. Based on feedback during pilot testing and from the data collection for national norms, the responses have been exceedingly positive with respect to the other test selection criteria, as well—the tests are considered “user-friendly” in terms of training, equipment, space, and time requirements. The participants generally report that the tests are enjoyable and motivating and that they look forward to receiving feedback about their performance and information on how to improve. Finally, the tests appear to be safe for the majority of the community-dwelling population of older adults. Among the more than 7,000 participants tested in the norms development phase of the project, no injuries or complications were reported.

Care should be taken, however, in generalizing the results of the reliability and validity studies reported in this article to other populations. The fact that the participants in these studies were volunteers, generally active, and well educated (see Table 1) may mean that the results would not apply to populations that are substantially different.

Normative Scores

The final stage of the test development is that of establishing normative scores for each of the test items. To assist in this process, a national study was conducted to collect test scores (along with other health and physical activity information) from over 7,000 adults (ages 60–90+) from 267 test sites in 21 states throughout the nation. The results, which include 5-year age-group norms for each test item for both men and women, are published in the following article in this journal.

In addition to developing normative performance standards, the ultimate goal will be that of identifying criterion-referenced standards—standards that will indicate the level of performance it is necessary to maintain in order to accomplish desired activity goals that might range from simple activities of daily living to
advanced-level activities including sports, travel, and recreational pursuits. Prior to
the development of such standards, however, additional research will be needed to
determine criterion-referenced reliability and validity.

References

and functional ability on performance biomechanics. Journal of Gerontology: Medi-
cal Sciences, 46, M91-M98.


American College of Sports Medicine. (1995). Guidelines for exercise testing and prescrip-

chronic diseases and disabilities. Champaign, IL: Human Kinetics.

physical activity for older adults. Medicine and Science in Sports and Exercise, 30,
992-1008.

American Psychological Association. (1985). Joint standards for education and psychologi-
cal tests. Washington, DC: author.

Leg extensor power and functional performance in very old men and women. Clinical
Science, 82, 321-327.


Bittner, V., Weiner, D.H., Yusuf, S., Rogers, W.J., McIntyre, K.M., Bangdiwala, S.I.,
Kronenberg, M.W., Kostis, J.B., Kohn, R.M., Guillette, M., Greenberg, B.,
a 6-minute walk test in patients with left ventricular dysfunction. JAMA, 270,
1702-1707.

als. In W.W. Spirduso & H.M. Eckert (Eds.), Physical activity and aging (pp. 120-130).
Champaign, IL: Human Kinetics.

muscles. Perceptual and Motor Skills, 80, 163-166.

Champaign, IL: Human Kinetics.

The functional fitness assessment battery: Reliability and validity data for elderly


Rubenstein, D. Wieland, & R. Bernabei (Eds.), Geriatric assessment technology: The
state of the art (pp. 147-159). Milan: Editrice Kurtis.


**Acknowledgment**

Appreciation is extended to the hundreds of people who participated in the numerous background studies for this project—especially to the test coordinators and participants at the Ruby Gerontology Center, Leisure World, Morningside Retirement Center, Fullerton Senior Center, and from the San Diego Adult Education programs. Special recognition also goes to the numerous students and faculty at California State University, Fullerton, who assisted with the data collection procedures.

We also extend our appreciation to all advisory panel members for their valuable contributions to this project. Local advisory panel members, all of whom are exercise specialists in southern California, were William C. Beam, Scott Duncan, Diane Edwards, Laura Gladwin, Blanche Lamar, Charlene Schade, and Greg Welch. National advisory panel members (with their affiliations at the time of service) were David M. Buchner, Departments of Health Services and Medicine, University of Washington; Wojtek Chodzko-Zajko, editor, *Journal of Aging and Physical Activity*, Kent State University; Janie Clark, president, American Senior Fitness Association, New Smyrna Beach, FL; Loretta DiPietro, Department of Epidemiology and Public Health, Yale University School of Medicine; Robert Dustman, director, neuropsychology research, V.A. Medical Center, Salt Lake City; William J. Evans, director, Noll Physiological Research Center, The Pennsylvania State University; Lawrence A. Golding, editor, *ACSM Health and Fitness Journal*, University of Nevada, Las Vegas; Sheldon Greenfield, Tufts University School of Medicine; Jack M. Guralnik, National Institute on Aging, Bethesda; William J. Haskell, professor of medicine, Stanford University; Vivian Heyward, professor of exercise science, University of New Mexico; Karl Knoph, president, Fitness Educators of Older Adults Association, Sunnyvale, CA; Jan Montague, director, Maple Knoll Village Wellness Center, Cincinnati, OH; James R. Morrow, Jr., chair of Kinesiology, Health Promotion, and Recreation, University of North Texas; Debra J. Rose, director, Motor Behavior Laboratory, Oregon State University; Kay Van Norman, program director, Young at Heart, Montana State University; and Shelley Whitlatch, Tucson Medical Center Fitness Center.

This work was supported by PacifiCare Health Systems.
Appendix A

FUNCTIONAL FITNESS TEST FOR OLDER ADULTS

Following are specific directions for administering each of the test items. To ensure scoring accuracy and interpretation, strict adherence to all instructions is essential. Throughout all testing, participants should be instructed to do the best they can on the tests but to never push themselves to a point of overexertion or beyond what they think is safe for them. Prior to testing, participants must do a 5- to 10-min warm-up and general stretching routine.

Based on guidelines established by the ACSM (1995) and on input from medical consultants, these tests are safe for the majority of community-residing older adults without medical screening and pose risks similar to those in engaging in other forms of moderate physical activity. People who should not take the tests without physician approval are those who

- Have been advised by their doctors not to exercise because of a medical condition
- Are currently experiencing chest pain, dizziness, or have exertional angina (chest tightness, pressure, pain, heaviness) during exercise
- Have had congestive heart failure
- Have uncontrolled high blood pressure (greater than 160/100)

30-Second Chair Stand

Purpose. To assess lower body strength.

Equipment. Stopwatch, straight-back or folding chair (without arms), height approximately 17 in. For safety purposes, the chair should be placed against a wall or in some other way stabilized to prevent it from moving during the test.

Protocol. The test begins with the participant seated in the middle of the chair, back straight and feet flat on the floor. Arms are crossed at the wrists and held against the chest. On the signal “go” the participant rises to a full stand and then returns to a fully seated position. The participant is encouraged to complete as many full stands as possible within 30 s. After a demonstration by the tester, a practice trial of one to three repetitions should be done to check for proper form, followed by one 30-s test trial.

Scoring. The score is the total number of stands executed correctly within 30 s. If the participant is more than half-way up at the end of 30 s, it counts as a full stand.

Arm Curl

Purpose. To assess upper body strength.

Equipment. Wristwatch with second hand, straight-back or folding chair (without arms), hand weights (dumbbells—5 lb for women, 8 lb for men).

Protocol. The participant is seated on a chair, back straight and feet flat on the floor, with the dominant side of the body close to the side edge of the chair. The weight is held at the side in the dominant hand (handshake grip). The test begins with the arm down beside the chair, perpendicular to the floor. At the signal “go” the participant turns the palm up while curling the arm through a full range of motion and then returns to the fully extended position. At the down position the weight should have returned to the handshake grip position.

The examiner kneels (or sits in a chair) next to the participant on the dominant-arm side, placing his or her fingers on the person’s mid-biceps to prevent the upper arm from moving and to ensure that a full curl is made (participant’s forearm should squeeze examiner’s fingers). It is important that the participant’s upper arm remain stabilized (still) throughout the test.

The examiner may also need to position his or her other hand behind the participant’s elbow so that the participant will know when full extension has been reached, as well as to prevent a back-swinging motion of the arm.

The participant is encouraged to execute as many curls as possible within the 30-s time limit. After a demonstration by the examiner, a practice trial of one or two repetitions should be given to check for proper form, followed by one 30-s trial.

Scoring. The score is the total number of curls made correctly within 30 s. If the arm is more than halfway up at the end of the 30 s, it counts as a curl.

6-Minute Walk

Purpose. To assess aerobic endurance.

Equipment. Stopwatch, long measuring tape, cones, popsicle sticks, chalk, masking tape (or some other type of marker). For safety purposes, chairs should be positioned at several points alongside the walkway.

Set-Up. The test involves assessing the maximum distance that can be walked in 6 min along a 50-yd course marked into 5-yd segments (see Figure A1). The inside perimeter of the measured distance should be marked with cones, and the
5-yd segments with masking tape or chalk. The walking area, which can be indoors or outdoors, should be well lit, with a nonslippery, level surface.

**Protocol.** To keep track of distance walked, a popsicle stick (or similar object) can be given to the participant each time he or she rounds a cone, or a partner can mark a score card each time a lap is completed. Two or more participants should be tested at a time, with starting times staggered (10 s apart) so that participants do not walk in clusters or pairs. When testing several people at once, numbers should be placed on the participants to indicate the order of starting and stopping. On the signal “go,” participants are instructed to walk as fast as possible (not run) around the course as many times as they can in 6 min. If necessary, participants may stop and rest (on provided chairs), then resume walking. The timer should move to the inside of the marked area after everyone has started. To assist with pacing, elapsed time should be called out when participants are approximately half done, when 2 min are left, and when 1 min is left. Encouragement phrases such as “you are doing well” and “keep up the good work” should be called out at approximately 30-s intervals. At the end of 6 min, participants (staggered every 10 s) are instructed to stop and move to the right, where an assistant will record their score. To assist with proper pacing and to improve scoring accuracy, a practice test should be given prior to the actual test day.

**Safety.** The test should be discontinued if at any time a participant shows signs of dizziness, pain, nausea, or undue fatigue. At the end of the test each participant should slowly walk around for about a minute to cool down.

**Scoring.** The score is the total number of yards walked in 6 min, to the nearest 5 yd. The test administrator or aide records the nearest 5-yd mark.

---

**2-Minute Step-in-Place (An Alternative to the 6-Min Walk Test)**

**Purpose.** To assess aerobic endurance.

**Equipment.** Stopwatch, tape measure or 30-in. piece of cord, masking tape, mechanical counter (if possible) to ensure accurate counting of steps.

**Set-Up.** The proper (minimum) knee-stepping height for each participant is at a level even with the midway point between the patella (middle of the knee cap) and the iliac crest (top hip bone). This point can be determined using a tape measure or by simply stretching a piece of cord from the patella to the iliac crest, then folding
it in half to determine the midway point. To monitor correct knee height when stepping, books can be stacked on an adjacent table or a ruler can be attached to a chair or wall with masking tape to mark the proper knee height.

**Protocol.** On the signal “go” the participant begins stepping (not running) in place, starting with the right leg, and completes as many steps as possible within the time period. Although both knees must be raised to the correct height to be counted, the tester only counts the number of times the right knee reaches it. The counter also serves as a spotter in case of loss of balance and ensures that the participant maintains proper knee height. As soon as proper knee height can no longer be maintained, the participant is asked to stop—or to stop and rest until proper form can be regained. Stepping may be resumed if the 2-min time period has not elapsed. If necessary, the participant can place one hand on the table or chair to assist in maintaining balance.

To assist with proper pacing and to improve scoring accuracy, a practice test should be given prior to the test day. On test day, the examiner should demonstrate the procedure and allow the participants to practice briefly to recheck their understanding of the protocol.

**Safety.** At the end of the test the participant should slowly walk around for about a minute to cool down.

**Scoring.** The score is the total number of times the right knee reaches the minimum height. To assist with pacing, participants should be told when 1 min has passed and when there are 30 s to go.

---

**Chair Sit-and-Reach**

**Purpose.** To assess lower body (primarily hamstring) flexibility.

**Equipment.** Straight-back or folding chair (approximately 17-in. seat height), 18-in. ruler. For safety purposes, the chair should be placed against a wall and checked to see that it remains stable (doesn’t tip forward) when the participant sits on the front edge.

**Protocol.** Starting in a sitting position on a chair, the participant moves forward until she or he is sitting on the front edge. The crease between the top of the leg and the buttocks should be even with the edge of the chair seat. Keeping one leg bent and foot flat on the floor, the other leg (the preferred leg*) is extended straight in front of the hip, with heel on floor and foot flexed (at approximately 90°; see the picture).

With the extended leg as straight as possible (but not hyperextended), the participant slowly bends forward at the hip joint (spine should remain as straight as possible, with head in line with spine, not tucked) sliding the hands (one on top of

---

*The preferred leg is defined as the one that results in the better score. Obviously, it is important to work on flexibility on both sides of the body, but for the sake of time, only the “better” side has been used in developing norms.
the other with the tips of the middle fingers even) down the extended leg in an attempt to touch the toes. The reach must be held for 2 s. If the extended knee starts to bend, ask the participant to slowly sit back until the knee is straight before scoring. Participants should be reminded to exhale as they bend forward; to avoid bouncing or rapid, forceful movements; and to never stretch to the point of pain.

After a demonstration by the tester, the participant is asked to determine the preferred leg. The participant is then given two practice (stretching) trials on that leg, followed by two test trials.

**Scoring.** Using an 18-in. ruler, the scorer records the number of inches a person is short of reaching the toe (minus score) or reaches beyond the toe (plus score). The middle of the toe at the end of the shoe represents a zero score. Record both test scores to the nearest 1/2 in., and circle the best score. The best score is used to evaluate performance. Be sure to indicate “minus” or “plus” on the score card.

---

**Back Scratch**

**Purpose.** To assess upper body (shoulder) flexibility.

**Equipment.** 18-in. ruler (half of a yardstick).

**Protocol.** In a standing position, the participant places the preferred hand* behind the same-side shoulder, palm toward back and fingers extended, reaching down the middle of the back as far as possible (elbow pointed up). The participant places the other hand behind the back, palm out, reaching up as far as possible in an attempt to touch or overlap the extended middle fingers of both hands.

Without moving the participant’s hands, the tester helps to see that the middle fingers of each hand are directed toward each other. The participant is not allowed to grab his or her fingers together and pull.

After a demonstration by the tester, the participant is asked to determine the preferred hand, and is then given two practice (stretching) trials, followed by two test trials.

*The preferred hand is defined as the one that results in the better score. Although it is important to work on flexibility on both sides of the body, only the “better” side has been used in developing norms.
**Scoring.** The distance of overlap or distance between the tips of the middle fingers is measured to the nearest 1/2 in. A minus score (−) is given to represent a distance short of touching; a plus score (+) represents the amount of an overlap. Record both test scores and circle the best one. The best score is used to evaluate performance. Be sure to indicate “minus” or “plus” on the score card.

**8-Foot Up-and-Go**

**Purpose.** To assess agility/dynamic balance.

**Equipment.** Stopwatch, tape measure, cone (or similar marker), straight-back or folding chair (seat height approximately 17 in.).

**Set-Up.** The chair should be positioned against a wall or in some other way secured so that it does not move during testing. It should also be in a clear, unobstructed area, facing a cone marker exactly 8 ft away (measured from a point on the floor even with the front edge of the chair to the back of the marker). There should be at least 4 ft of clearance beyond the cone to allow ample turning room for the participant.

**Protocol.** The test begins with the participant fully seated in the chair (erect posture), hands on thighs and feet flat on the floor (one foot slightly in front of the other). On the signal “go” the participant gets up from the chair (pushing off thighs or chair is allowed), walks as quickly as possible around the cone (on either side), and returns to the chair. The participant should be told that this is a timed test and that the object is to walk as quickly as possible (without running) around the cone and back to the chair. The tester should serve as a spotter, standing midway between the chair and the cone, ready to assist the participant in case of loss of balance. For reliable scoring, the tester must start the timer on “go,” whether or not the participant has started to move, and stop the timer at the exact instant the participant sits in the chair.

After a demonstration, the participant walks through the test one time as a practice and then is given two test trials. Participants should be reminded that the timing does not stop until they are fully seated in the chair.

**Scoring.** The score is the time elapsed from the signal “go” until the participant returns to a seated position in the chair. Record both test scores to the nearest 1/10th s and circle the best score (lowest time). The best score is used to evaluate performance.
## Appendix B

GUIDELINES FOR CONVERTING NONMETRIC MEASURES TO METRIC UNITS

<table>
<thead>
<tr>
<th>Test item</th>
<th>U.S. measure</th>
<th>Metric equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair stand</td>
<td>17-in. seat height</td>
<td>43.18-cm seat height</td>
</tr>
<tr>
<td>Arm curl</td>
<td>5-lb dumbbell</td>
<td>2.27-kg dumbbell</td>
</tr>
<tr>
<td></td>
<td>8-lb dumbbell</td>
<td>3.63-kg dumbbell</td>
</tr>
<tr>
<td>Chair sit-and-reach</td>
<td>17-in. seat height</td>
<td>43.18-cm seat height</td>
</tr>
<tr>
<td></td>
<td>18-in. ruler</td>
<td>45.72-cm ruler</td>
</tr>
<tr>
<td></td>
<td>scores recorded to nearest 1/2 in.</td>
<td>scores recorded to nearest cm</td>
</tr>
<tr>
<td>Back scratch</td>
<td>18-in. ruler</td>
<td>45.72-cm ruler</td>
</tr>
<tr>
<td></td>
<td>scores recorded to nearest 1/2 in.</td>
<td>scores recorded to nearest cm</td>
</tr>
<tr>
<td>6-min walk</td>
<td>50-yd course marked into 5-yd segments</td>
<td>45.72-m course marked into 4.57-m segments</td>
</tr>
<tr>
<td>2-min step</td>
<td>30-in. cord or tape measure</td>
<td>76.2-cm cord or tape measure</td>
</tr>
<tr>
<td>8-ft up-and-go</td>
<td>17-in. seat height</td>
<td>43.18-cm seat height</td>
</tr>
<tr>
<td></td>
<td>8-ft distance to cone</td>
<td>2.44-m distance to cone</td>
</tr>
<tr>
<td></td>
<td>4-ft clearance past cone</td>
<td>1.22-m clearance past cone</td>
</tr>
</tbody>
</table>