Assessing Physical Performance in Independent Older Adults: Issues and Guidelines

Roberta E. Rikli and C. Jessie Jones

With the projected growth in the older adult population, preventing or delaying physical disability in later years has become a national goal. Evidence suggests that physiological decline, especially that associated with physical inactivity, is modifiable through proper assessment and activity intervention. However, a major limitation in reducing loss of function in later years is the lack of suitable assessment tools. Especially lacking are tests that can measure physical performance on a continuum across the wide range of functioning in the independent, community-residing older adult population. Of special concern is the ability to assess underlying physical parameters associated with common activities of daily living. Additional tools are needed for measuring physical performance in older adults, especially tools that meet established guidelines in terms of reliability, validity, discrimination power, and performance evaluation standards.

Key Words: aging, assessment, mobility, functional ability

Physical frailty in later life is costly in terms of both money spent on medical care and the diminished quality of life that comes from years of chronic disabilities. The estimated annual cost of physical frailty is $54–80 billion and could grow to over $132 billion by the year 2030 unless disability rates are lowered (Select Committee on Aging, 1992). With the surging growth of the aging population, especially among the oldest old (85+ years) (Tauber & Rosenwaike, 1992), preventing or delaying physical frailty and increasing the number of healthy years for older people are major goals for the nation, according to the Healthy People 2000 initiative (Department of Health and Human Services, 1990). Statistics in the Healthy People 2000 report indicate that in 1981, Americans spent on average 11.7 years (nearly 16% of their lives) with chronic disabilities that limited activities of daily living. Unfortunately, the recently completed Healthy People 2000: Midcourse Review and 1995 Revisions (Department of Health and Human Services, 1996a) indicates that the nation may be “losing ground” on this important goal; as life expectancy for Americans continues to increase, so too does the possibility of living more years with major physical limitations.

Evidence suggests that as much as 50% of the age-related decline leading to frailty may be preventable through early detection of weaknesses and proper

Roberta E. Rikli and C. Jessie Jones are with the Division of Kinesiology and Health Promotion, California State University, Fullerton, CA 92834.
adjustments in physical activity behaviors (Jackson, Beard, Wier, & Blair, 1995; Smith, 1980). However, one of the main problems associated with understanding and reducing loss of function in later years is the inability to adequately measure physical performance in older adults (Chodzko-Zajko, 1994; Guralnik, Branch, Cummings, & Curb, 1989; Spirduso, 1995). The purpose of this paper is to discuss the importance of assessing physical performance in older individuals, present a brief overview of physical assessments for people of advanced age, and provide some guidelines for future test development, including a discussion of special problems associated with assessing elderly people.

**Importance of Assessing Physical Performance in Older Adults**

Most people would agree that, to a large degree, quality of life is defined by the ability to do what one wants to do, without pain, for as long as possible. In fact, loss of physical function and physical independence is one of the greatest fears of older adults (Atchley, 1988). Preventing or reducing loss of function in later years depends, in part, on the ability to detect and treat any physical declines that may be precursors to more serious loss of function. Many older adults, often due to their sedentary lifestyles, function dangerously close to their maximum ability level during normal activities of daily living. Climbing stairs or getting out of a chair, for example, requires near-maximum effort for many older individuals (Evans, 1995). Any further decline or small physical setback could easily cause additional physical limitation that leads to disability and dependency. The unfortunate consequence is that a physical weakness that possibly could have been remedied through low-cost evaluation and exercise now may require costly rehabilitation or long-term health care.

A number of measures have been developed to identify functional limitations in the dependent and frail elderly, such as the ADL (Katz, Ford, & Moskowitz, 1963) and IADL (Lawton & Brody, 1969) scales, which assess basic and instrumental activities of daily living (e.g., bathing, dressing, preparing meals, and shopping). However, few tests have been designed to evaluate performance in the large proportion of older adults who reside independently in the community. Especially lacking are instruments capable of assessing declines in the underlying physical parameters (e.g., muscle function, aerobic endurance, and flexibility) that lead to loss of function (Spirduso, 1995). The functional performance framework outlined in Figure 1 indicates a progressive relationship between physiological performance, functional performance, and activity goals.

As indicated in the far right column of Figure 1, common activity goals (e.g., personal care, shopping, and traveling) require the ability to perform the functions listed in the middle column (e.g., walking and stair climbing). Functions such as walking and stair climbing, in turn, require physical strength, endurance, flexibility, and motor ability as identified in the column on the left. Certainly, detecting and treating physical impairments (declining strength, endurance, etc.) should be a critical step in preventing or slowing the progression toward functional limitation, disability, and dependency.

UNDERSTANDING AND EVALUATING PHYSICAL CHANGE

Developing tools for assessing physical function, as well as planning effective intervention strategies, requires an understanding of the progressive stages of
FUNCTIONAL PERFORMANCE FRAMEWORK

<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></td>
<td>Shopping/errands</td>
</tr>
<tr>
<td>Aerobic endurance</td>
<td>Stair climbing</td>
<td>Housework</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Standing up from chair</td>
<td>Gardening</td>
</tr>
<tr>
<td>Motor ability</td>
<td>Lifting/reaching</td>
<td>Sports</td>
</tr>
<tr>
<td>balance</td>
<td>Bending/kneeling</td>
<td>Traveling</td>
</tr>
<tr>
<td>coordination</td>
<td>Jogging/Running</td>
<td></td>
</tr>
<tr>
<td>speed/agility power</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Physical impairment | Functional limitation | Physical disability/dependence |

Figure 1. A physical performance framework demonstrating the progressive relationship between physiological performance, functional performance, and activity goals.

change leading to frailty. Traditional medical models describing the progression to disability, such as those proposed by Nagi (Nagi, 1991) and the World Health Organization (see Berg & Cassells, 1990), suggest that disability is preceded by disease or chronic conditions. The progressive steps, as described by Nagi (1991), indicate that pathology generally leads to impairment, impairment progresses to functional limitation, and functional limitation leads to disability (see Figure 2a). However, evidence now indicates that as more people are living into their 80s and 90s, a physically inactive lifestyle can also be a primary cause of dysfunction in later years, suggesting that the model for explaining disability (especially physical disability) should be modified as indicated in Figure 2b. In the amended model, the directional arrows suggest that physical inactivity (disuse) and pathology each have independent as well as interrelated effects on the processes leading to disability.

Physical inactivity has an independent or direct effect on physical impairment and functional decline due to the considerable physical atrophy that results from disuse. Some of the early and most dramatic evidence demonstrating the impact of inactivity came from bed rest and immobilization studies. Even in healthy young people, serious muscle atrophy and marked declines in physiological functioning (sometimes equivalent to as much as 20 years worth of aging) were observed following only a few weeks of immobility (Bloomfield & Coyle, 1993; Bortz, 1982; Saltin et al., 1968; Shephard, 1994). Of more universal concern, however, is that for people with generally sedentary lifestyles, the strength and endurance thresholds needed for independent living are commonly reached by about the age of 80 or shortly thereafter (Shephard, 1993). Additional support for the role of physical inactivity relative to disability comes from recent epidemiology studies on factors affecting physical functioning in older adults (DiPietro, 1996). Data from these studies (Kaplan, Strawbridge, Camacho, & Cohen, 1993; Mor et al., 1989; Seeman
Figure 2. (a) Nagi's (1991) model suggesting that the progression to disability originates with disease/pathology, and (b) an amended version of the model indicating that disuse, as well as disease, can lead to disability.

et al., 1995) indicate that physical inactivity is on a par with chronic disease as a determinant of functional decline (DiPietro, 1996). Results from the Established Populations for Epidemiological Studies (EPESE) (Lacroix, Guralnik, Berkman, Wallace, & Satterfield, 1993) and from the Medical Outcomes Study (Stewart et al., 1994) provide still further support for the effects of physical activity, independent of disease, on functional capacity. Findings from these studies show that physical activity participation is associated with higher levels of mobility, even for older adults who already have chronic conditions.

Physical inactivity also has an indirect impact on the progression toward disability because of its influence on the disease process. The association between inactivity and the major chronic diseases of aging (coronary heart disease, hypertension, stroke, non-insulin-dependent diabetes, osteoporosis, colon cancer, and some mental health problems) has been well documented in scientific reviews (Bouchard, Shephard, & Stephens, 1994), by leading health organizations (Pate et al., 1995), and most recently in the Surgeon General's report on physical activity and health (Department of Health and Human Services, 1996b).

Most relevant to our discussion of physical testing, however, is the indication that physical decline, whether due to disease or disuse, is modifiable through proper assessment and activity intervention. A number of research studies provide evidence that physical activity, even when begun late in life, is associated with improved physiological measures and functional performance, regardless of disease status (Cress et al., 1991; Fiatarone et al., 1990, 1994; Nichols, Hitzelberger, Sherman, & Patterson, 1995; Pyka, Lindenberger, Charette, & Marcus, 1994; Rikli & Edwards, 1991). In these studies, where the incidence of chronic conditions was similar for both exercising and control subjects, the exercising subjects demonstrated significantly higher levels of strength, endurance, and motor function compared to nonactive subjects.

Whatever the cause of physical impairment, proper assessment can help to identify the specific deficits in physical parameters that will need attention if further
decline is to be prevented or reduced. Specifically, the ability to assess physical performance in older adults can increase the quality of feedback provided to individuals, help program developers to plan more effective intervention programs, provide better outcome measures for evaluating both individual improvement and program effectiveness, and enhance understanding of age-related declines. In addition, personal assessment helps to empower older adults by providing them with information about their physical ability levels and about activities they can undertake to improve functioning. Furthermore, it has been our experience that most older people enjoy the attention and feedback associated with personalized assessments and, due to a heightened awareness of their physical ability levels, are motivated to set goals for maintaining or improving their performance on subsequent evaluations.

Physical Assessments for People of Advanced Age

Increasingly, professionals in gerontology and exercise science are emphasizing the need for further test development to assess physical parameters, particularly tests that can measure across the wide ranges of older adults’ physical abilities (Chodzko-Zajko, 1994; Haskell & Phillips, 1995; Spirduso, 1995). To understand the physical evaluation needs concerning people of advanced age, it is important to consider the various levels of functional status. Spirduso (1995, pp. 338–355) identified the following five hierarchical categories to differentiate among the common levels of functioning:

1. The physically dependent, the lowest level of the hierarchy, cannot execute basic activities of daily living (ADLs) (dressing, bathing, eating, etc.) and, therefore, are dependent upon others for their daily functioning. Individuals in this category are institutionalized or require full-time care at home.

2. The physically frail can perform all ADLs but not all instrumental activities of daily living (IADLs) (i.e., cannot do household activities such as preparing meals, shopping, and housekeeping). These individuals may still live at home but require outside services such as Meals on Wheels or home nursing care.

3. The physically independent can do all IADLs—they are fully independent at home and within the community—but generally are quite sedentary. People within this group have wide variations in ability, from those who function at a high level to those who are borderline frail and close to losing their ability to function independently.

4. The physically fit exercise regularly, usually several times a week. They typically have well above average functional ability and often appear to be younger than their chronological age. Many participate in strenuous activities such as jogging, aerobic exercise classes, tennis, and swimming.

5. The physically elite are the very small percentage of older adults who regularly train for competitive activities. They may be involved in master athlete competitions or may participate in community events such as races.

Obviously, the type of tests required to assess physical ability in each of these groups would vary considerably, depending on the purpose of testing and the subjects’ physical and cognitive capabilities. The majority of test development for
older adults has been directed toward the first two categories, the physically
dependent and the physically frail, who constitute approximately 25% of the elderly
population (Fitti & Kovar, 1988). Various reviews (Feinstein, Josephy, & Wells,
1986; Spirduso, 1995) refer to more than 50 published scales that have been
developed for these lower ability groups, primarily to assess functional limitations
in order to provide services and/or rehabilitation. Most of these tests use either a
simple yes/no self-evaluation scale or observer rating scale to assess functional
abilities such as personal care, balance, gait, and stair climbing (Heath, 1989; Jette,
Branch, & Berlin, 1990; Reuben, Siu, & Kimpau, 1992). Since few of these tests can
measure physical performance on a continuum, they contribute little to our ability
to study changes in function over time, particularly changes in higher levels of
functioning that may precede disability. Most individuals in the physically indepen-
dent category (approximately 70% of adults over 75) and those few (approximately
5%) in the physically fit and physically elite categories (Spirduso, 1995) would
most likely receive perfect scores on all of these test items.

Although the traditional ways of assessing fitness in younger populations
(e.g., treadmill or cycle endurance tests or 1 repetition maximum [1 RM] strength
tests) may be viable for use with high-functioning older people (the fit and elite),
they generally are not suitable for the majority of the less active older adults. Many
such protocols either are too difficult for elderly people or require costly equipment
and/or close monitoring of medical conditions, neither of which is feasible in most
field settings where older adults are tested. As suggested by Spirduso (1995), the
large physically independent category of individuals has “fallen between the
cracks” with respect to our ability to evaluate and understand their physical
capabilities. While the frail and disabled can be assessed through the numerous
scales referred to previously, and the fit and the elite through the many protocols
developed for younger people, there are few standardized tests designed to evaluate
physical performance in the 70% who are in the middle category.

Progress is being made in the development of new physical performance test
batteries for older adults (Kim & Tanaka, 1995; Kimura, Hirakawa, & Morimoto,
1990; Kinugasa et al., 1994; Lemminck et al., 1995; Oness et al., 1990), but much
of this work has not yet been widely published. The exception, at least within the
United States, is the frequently cited American Alliance for Health, Physical
Education, Recreation, and Dance (AAHPERD) Functional Fitness Assessment for
Adults Over 60 Years (Oness et al., 1990), which measures muscular strength/endurance, coordination, agility/dynamic balance, flexibility, and aerobic endurance.
The test items, which were designed to be used in a community setting, are easy to administer, safe for participants, and low in cost.

Although the AAHPERD assessment battery has made a landmark contribu-
tion to the functional assessment literature for older adults, certain adjustments
would increase its effectiveness considerably. For example, the assessment does not
have a lower body strength item. Since a decline in lower body function has been
identified as a major predictor of subsequent disability (Guralnik, Ferrucci, Simonsick,
Salive, & Wallace, 1995; Guralnik et al., 1994; Phillips & Haskell, 1995) and is
strongly associated with incidence of falling (Lord, McLean, & Stathers, 1992;
Whipple, Wolfson, & Amerman, 1987), adding such a test item is highly recom-
mended. Also, the AAHPERD test of aerobic endurance, a half-mile walk, mea-
ures only those who can walk that distance, thus eliminating much of the older adult
population from testing. Statistics indicate that 40% of community-residing older adults over 75 have difficulty walking even one-fourth of a mile (Select Committee on Aging, 1992). Modifying the test so that a larger proportion of the target population could be assessed would greatly improve its usefulness. As suggested later in this paper (see Discrimination Power), changing the test protocol to a prescribed time (such as 9 min) instead of a set distance would make it possible to assess all ability levels.

Examples of other test items designed to assess physical performance in independent older adults are those included in the National Institute of Aging EPESE projects (Guralnik et al., 1994, 1995) and in the MacArthur Successful Aging studies (Seeman et al., 1994). Although the measures used in these studies meet the intended purpose (i.e., predicting risk of disability for specific groups of people), they do not, in some cases, discriminate well across the full range of individual functioning. For example, a chair stand test used to assess lower body strength was found to be too difficult for a sizable portion of the target population; over 20% of the participants were unable to complete the test (Guralnik et al., 1994). A 10-s tandem balance test, on the other hand, was too easy to discriminate among many of the individuals, with nearly half of the subjects obtaining perfect scores (Guralnik et al., 1994; Seeman et al., 1994). Tests are needed that can measure performance on a continuum across all levels of ability so that baseline measures can be established and rate of change can be accurately assessed for the entire population. (See section on Discrimination Power for further discussion.)

The recent interest in assessing higher level functioning in older adults also led to development of the Advanced Activities of Daily Living (AADL) scale (Reuben, Laliberte, Hiris, & Mor, 1990). Although this scale has been successful in identifying exercise behavior categories (i.e., frequent vigorous exercisers, frequent long walkers, frequent short walkers, and nonexercisers), it was not designed to measure actual physical performance parameters. It is clear that additional tools are needed to adequately assess the continuum of physical performance in older adults. The remaining sections of this paper will suggest guidelines for future test development.

**Guidelines for Future Test Development**

Assessment protocols for older adults, whether for research or for individual evaluation, should meet established test construction guidelines with respect to reliability, validity, discrimination power, and performance standards (American Psychological Association [APA], 1985; Safrit & Wood, 1995). Test items also should be socially acceptable to older participants and, preferably, feasible for use in the field setting. The following are special considerations to be addressed when selecting or developing such assessments.

**RELIABILITY AND VALIDITY**

Test reliability and validity, the essential qualities of any test, must be documented using age-appropriate subjects. It can never be assumed that tests developed and validated for a younger age group will be appropriate for older people. The changing characteristics and increased heterogeneity commonly observed in older adults with
respect to health status and physical and mental ability levels (Evans & Rosenberg, 1991) often make it necessary to adjust test protocols.

**Reliability.** Test reliability refers to the dependability of test scores. It reflects the degree of consistency of an individual (or group of individuals) in performing a test (Safrit & Wood, 1995). Measures that are consistent from one trial to the next (especially from one day to another) are said to have stability reliability, an especially important characteristic if performance is to be evaluated over time. Establishing stable (reliable) baseline measures is critical to the accurate evaluation of program or experimental intervention effectiveness. A recommended method of estimating the stability reliability of physical performance tests is to correlate test–retest scores given on two different days (usually two or three days apart), preferably by calculating the intraclass coefficient \( R \) using analysis of variance (ANOVA) (Baumgartner & Jackson, 1995; Safrit & Wood, 1995). The intraclass analysis, as opposed to the interclass Pearson product moment procedure (\( r \)), reflects not only the relative relationship of test–retest scores (as does the interclass method) but also provides information on any systematic increase or decrease in day-to-day scores.

The recommended steps in computing intraclass \( R \) are to obtain an ANOVA summary table and check the corresponding \( F \) value for significant trial effects. If \( F \) is not significant, \( R \) is computed using an appropriate formula (Baumgartner & Jackson, 1995). A high \( R \) value, combined with a nonsignificant trial effect, would indicate a stable (reliable) measure. If the \( F \) value for trials is significant, testing protocols should be modified and the process repeated until a stable measure can be identified. Tests, especially when used as baseline measures, should have high test–retest correlations (relative reliability) as well as nonsignificant trial-to-trial changes in scores. Generally, test reliability should be determined using sample sizes of at least 30, and preferably 50 in situations where the resulting reliability coefficient is apt to be less than .90 (Morrow, Jackson, Disch, & Mood, 1995). Although there is no agreed upon minimum reliability coefficient, values generally should be .80 or higher, with estimates below .70 rarely considered acceptable (Morrow et al., 1995; Safrit & Wood, 1995). Additional information on methods of estimating reliability in physical performance tasks can be found in most measurement textbooks (Baumgartner & Jackson, 1995; Morrow et al., 1995; Safrit & Wood, 1995).

A variety of factors can affect the potential reliability of a test. Tests that measure maximum effort, such as strength tests, usually have higher reliability than tests which measure accuracy or skill level. Scoring instability (unreliability) also can result from fluctuations in individual attributes or behaviors. Performance variations in older adults might be caused by a number of conditions such as fluctuations in fatigue or pain level, reduced information processing speed, changing medical conditions or medications, and task (or equipment) unfamiliarity.

Individual scoring variability can be represented statistically by the standard error of measurement (SEM). Theoretically, SEM is calculated by observing a subject’s performance on multiple occasions and then computing the standard deviation of the individual trials about the mean of the trials (Safrit & Wood, 1995). When an individual’s scores are relatively consistent across trials, the standard deviation (and thus the \( SEM \)) is small. However, if there is considerable variation from score to score, the standard deviation (and thus the \( SEM \)) will be large. Small SEM values indicate stable or reliable measures, whereas a large SEM reflects scoring inconsistency and lack of reliability. When it is not possible or practical to
observe a person's performance on numerous occasions (usually it is not), a formula can be used to estimate SEM (Safrit & Wood, 1995).

Although few data are available on SEM differences across age groups, there is some indication that scoring variability increases with age on certain types of motor and psychomotor variables but remains constant on other, more physical types of tasks. In a single study (Rikli & Busch, 1986) comparing 30 young (ages 18–31) and 30 older (ages 61–80) subjects, SEM values were considerably higher for the older subjects on a simple reaction time task ($SEM = 26.8$ ms for older subjects vs. $7.4$ ms for younger subjects) and on a one-leg balance test ($SEM = 9.2$ s for older subjects vs. $0.0$ s for younger subjects). The lack of scoring variability for younger subjects on the balance task was due to the fact that they all obtained the maximum score of 60 s. In the same study, however, the scoring variability on physical (as opposed to psychomotor) tasks involving trunk flexibility, shoulder flexibility, and grip strength was nearly identical between age groups, even though older subjects' overall performance was lower.

Increasing scoring consistency, and therefore reliability, on certain tests for older adults may require adjustments in testing protocols. Providing additional practice, longer rest periods, clearer instructions, and/or an increased number of testing trials might improve scoring consistency. In one study, for example, it was determined that older adults required nearly four times the number of trials to achieve a stable criterion measure of simple reaction time than did younger subjects. Whereas only 8 trials were needed to achieve reliable measures for subjects ages 18 to 40 ($R = .87$) (Haywood & Teeple, 1976), 30 trials were required to achieve acceptable consistency ($R = .83$) for adults ages 40 to 67 (Lupinacci, Rikli, Jones, & Ross, 1993). In another study involving strength training, additional practice time and testing trials were required to achieve reliable 1 RM strength measures for older adults who had little experience with resistance training equipment (Rikli, Jones, Beam, Duncan, & Lamar, 1996).

Test objectivity, a form of reliability (sometimes called rater reliability), reflects the degree of accuracy in scoring a test. Scoring accuracy is an especially important test characteristic if measures are to be collected on multiple occasions over extended time intervals either by the same evaluator (intrarater reliability) or by different evaluators (interrater reliability). Both types of scoring reliability can be computed using either the Pearson product moment correlation coefficient ($r$) or the intraclass correlation coefficient ($R$) (Safrit & Wood, 1995).

Test objectivity (scoring accuracy) can be maximized by following clearly defined scoring instructions. On a short timed test, for example, it would be critical that all examiners follow identical procedures in operating a stopwatch. Considerable scoring error will result if one examiner starts a stopwatch on the signal “go” while another starts it when a participant actually begins to move. Both intrar- and interrater agreement can be improved through use of a carefully prepared script describing test directions as well as through adequate training and practice prior to each test administration. A well-written testing script not only provides the testers with clear prompts regarding the testing protocol but also ensures that the test takers receive clear and consistent instructions regarding test procedures.

Validity. A valid test is one that measures what it is intended to measure (Safrit & Wood, 1995). Therefore, in evaluating the validity of a test, it is important to consider the purpose of the test in addition to its stability or consistency. A test
may be valid in one setting and not in another. For example, a particular health-related fitness test may be valid for young adults whose primary goal is health maintenance but not valid for older adults whose main concern is functional mobility. For older adults, many of whom already have chronic conditions, the focus tends to shift from disease prevention to disability prevention (Berg & Cassells, 1990). When selecting or evaluating tests to assess physical performance, researchers should consider three types of validity: content (or logical) validity, concurrent validity, and construct validity (APA, 1985).

Content validity (or logical validity) is the degree to which a test (or test battery) reflects a defined “universe” of content (APA, 1985). A first step in ensuring the content validity of a test is to identify, typically through a literature review, an expert panel, and/or factor analysis, the important components of the construct (domain) of interest. Assuming that the physical parameters discussed earlier (listed in the first column of Figure 1) have been determined through research and/or expert judgment to be the underlying components of functional capacity, then a content-valid fitness test battery for older adults would be one that includes measures of muscular function (lower and upper body), aerobic endurance, and flexibility (lower and upper body) and one or more motor ability measures that reflect balance, coordination, agility, speed, and power. The use of factor analysis to identify the physical and motor dimensions of functional capacity was demonstrated in a recent article by Greene, Williams, Macera, and Carter (1993).

Concurrent validity represents the degree to which a test correlates with a criterion measure (i.e., a measure that is already known to be valid). Directly measured VO₂max, for example, traditionally has been the accepted criterion measure or gold standard of physical work capacity against which all other types of aerobic capacity tests are compared (validated). A field test of aerobic endurance (such as a submaximal distance walk or step test) could be validated by comparing performance on it with performance on the criterion VO₂max test. Concurrent validity of the field test, typically assessed through Pearson product moment correlation analysis with the criterion measure, is represented by the statistic r. If r is high enough, preferably above .80 (Safrit & Wood, 1995), then the substitute (field) test is considered a valid estimate of the criterion measure. Similarly, in strength testing, 1 RM is often considered the gold standard for validating submaximal strength measures. In theory, any new physical performance test should be validated by comparing scores with a proven measure of that test component.

However, selecting the appropriate criterion measure against which to validate fitness tests for older adults may require different logic than for younger people. Whereas the measurement of aerobic capacity (VO₂max) is considered to be the single best indicator of overall fitness for youth and young adults, gerontology researchers question its relative importance for older subjects whose critical goal is maintaining mobility (Phillips & Haskell, 1995). In later years, muscular fitness (especially lower body strength) generally is considered more important to daily functioning and quality of life than is cardiovascular fitness (Manton & Soldo, 1992; Pendergast, Fisher, & Calkins, 1993).

Also, maximum oxygen uptake, due to its strong genetic component, particularly in the surviving, genetically endowed population of older adults, appears to lose some of its sensitivity to training (environmental) effects. As a marker of
fitness, $\text{VO}_2\text{max}$ in older adults may indicate the effects of heredity more than the effects of exercise (DiPietro & Seals, 1995). Endurance training with older subjects, for example, has been found to result in as much as 400–500% improvement in functional ability but in negligible (10–15%) improvement in $\text{VO}_2\text{max}$ (Haskell & Phillips, 1995). It has been suggested that for older adults, traditional indices of fitness (e.g., $\text{VO}_2\text{max}$) probably should be “superseded, or at least complemented” by tests that more accurately reflect the functions required for activities of daily living (Haskell & Phillips, 1995). Tasks are needed, for example, that reflect one’s ability to perform common activities of daily living such as rising from a chair, walking, stair climbing, lifting, reaching, and bending.

Because of the difficulty in identifying any one acceptable standard of fitness for older adults, and because of the multidimensionality of functional capacity in later years (Greene et al., 1993), it seems necessary to use multiple indicators to determine a person’s level of fitness. Similarly, to estimate the criterion-referenced validity of older adult fitness test items, it may be desirable to consider a combination or composite of criterion laboratory tests (tests involving endurance, strength, range of motion, coordination, etc.) as the standards of comparison.

Construct validity is the degree to which a test measures a particular construct of interest. A construct is an attribute that exists in theory but cannot be directly observed, such as intelligence, personality, or in this case, functional fitness. Some measurement experts (Morrow et al., 1995) consider construct-related validity to be the highest form of validity since it combines both theoretical (logical) and statistical evidence in establishing validity, thus providing additional support for content and criterion-referenced validity as well. Determining construct validity involves theorizing that if a particular construct is valid, certain predictable relationships should exist. For example, if a test is a valid measure of physical fitness, then observable differences should exist between groups known to have different levels of fitness. Test scores for physically active older adults should be higher than those for physically independent but sedentary individuals. Similarly, independent sedentary older adults should score higher than those who have been identified as physically frail.

The procedure just described for determining construct validity is called the “known groups” or “group differences” method (Morrow et al., 1995). Differences between groups are tested for significance using a $t$ test (for two groups) or ANOVA (for more than two groups). A minimum significance level of .01 is suggested to increase the probability that group differences are real and not a result of chance (Baumgartner & Jackson, 1995).

**DISCRIMINATION POWER**

The amount of discrimination power needed for a test depends on the purpose of the test and the variability of the individuals for whom the test is designed. If a goal is to establish baseline measures in order to assess change over time, or in relation to a particular intervention, then the test will need to measure a continuum of performance (without “floor” or “ceiling” effects) across wide ranges of ability—from the borderline frail to the high functioning. A floor effect occurs when a test is too difficult for a large portion of the intended subjects. The AAHPERD half-mile walk test has a floor effect for many older people: Anyone who is unable to walk
half a mile cannot reach the floor (minimum requirements) of the test and cannot, therefore, receive a score.

A ceiling effect, on the other hand, occurs when a test is too easy for much of the population of interest. If a significant proportion of a subject group reaches a perfect score on a test item, then that test’s ceiling is too low to discriminate among (assess) this portion of the population. Such a condition occurred on the 10-s tandem balance tests in the EPESE and MacArthur Successful Aging studies, when 40% or more of the target populations received perfect scores (Guralnik et al., 1994; Seeman et al., 1994).

In some cases, a simple adjustment in testing protocol can eliminate a potential floor or ceiling effect, thus markedly improving the discrimination power of the test. For example, setting an amount of time (such as 9 min) as the goal on a walking test instead of a distance (such as half a mile) would make it possible to assess all levels of walkers ranging from the most frail to the most fit. On a timed walking test, scores can be obtained for those who can walk only a few feet as well as for those who can cover several hundred yards in the prescribed time. Similarly, standardizing the time (such as 30 s) instead of the number of required repetitions on the chair stand test reported in the Guralnik et al. study (1994) would greatly improve the test’s range and discrimination ability. On a timed test, those who are able to complete only one stand (or no stands) in 30 s can be scored, as can those who are able to perform numerous stands in the allotted time. In the Guralnik et al. study, continuous scoring was not possible for the 22% of the study population who could not complete the minimum requirement of five stands.

Although it may not be possible to design one assessment tool that can differentiate across all levels of functional hierarchy in older adults, it seems important to be able to differentiate across the full range of functioning within independent, community-residing elderly and into the “edge” of the frail category so that changes from one level to another can be monitored. It is important not only to detect weaknesses in the physically independent so that adjustments can be made to prevent further decline leading to frailty, but also to identify areas of weakness in the frail elderly which, if strengthened, might restore their independent functioning.

**PERFORMANCE STANDARDS**

Another important part of test development is establishing standards for evaluating performance. Performance can be evaluated relative to a peer group (norm-referenced standards) or in relation to predetermined, desired outcomes (criterion-referenced standards). Norm-referenced standards, generally reported in percentile or T-score tables, represent typical performances of a defined population, such as women over the age of 60. The construction of normative tables requires collecting test scores on large numbers of participants, preferably several hundreds of individuals within each subgroup category (Safril & Wood, 1995). Ideally, physical performance norms for older adults should be categorized by gender and age, utilizing no larger than 5-year age intervals (i.e., 60–64, 65–69, etc.) (Spirduso, 1995). Referring to normative standards makes it possible to assess an individual’s performance with respect to peer group.
Criterion-referenced standards allow researchers to evaluate performance relative to what is needed or recommended in order to achieve a particular level of health or function, regardless of how other people score. Developing criterion standards is more complex than developing norms, requiring that a particular level of performance be referenced to a criterion behavior or characteristic. In a health-related fitness test, for example, a criterion standard for aerobic capacity might be that level which research has found to be associated with reduced incidence of cardiovascular disease or other health conditions. If assessing functional mobility is the objective, a fitness criterion might be set according to the level of performance a person needs to function independently within the community or to perform some other activity, such as playing golf, lifting a grandchild, or carrying a suitcase. Determining the walking speed (feet per second) needed to cross an intersection within the average time allowed by street lights is an example of a criterion-referenced standard for a walking speed test. Certainly, the development of other performance standards, such as the amount of strength and endurance needed to carry out specific activities, would be a major contribution to the assessment literature for older adults.

In the early stages of test development, criterion standards often are based largely on normative data combined with the best judgment of experts in the field (Morrow et al., 1995). Ratings of “good” and “excellent,” for example, initially may be determined by what has been observed to be “above average” and “far above average” for a particular group of people. As more studies are conducted using the test items, standards can be refined and better referenced to specific activity or health goals.

SOCIAL ACCEPTABILITY

For optimum motivation, a test should be meaningful and acceptable to the participant. The social acceptability of a test may vary according to a person’s age, gender, economic level, and ethnic and cultural background. If a particular activity has a negative connotation within a certain culture, then its people most likely would not find a test involving that activity to be very meaningful. In the Navajo culture, for example, where walking is viewed largely as a means of transportation for those who cannot afford an automobile, a test involving walking may not be considered very positive or motivating (Arviso, 1996). In such cases, either an alternative type of test should be administered, or the people may need to be educated as to alternative values of the activity.

FEASIBILITY

The feasibility of a test has to do with its suitability for use and probability of being used in a particular testing environment. Cost and inconvenience generally prohibit transporting large numbers of older adults to laboratories for assessment. Therefore, most physical assessment, whether for individual evaluation, program evaluation, or research (particularly large-scale research projects), is likely to take place in a community setting. Tests suitable for use in the field should be relatively easy to administer and score, should be safe for participants (without medical supervision), and should require minimum equipment, time, and space.

The lack of available and easy-to-administer field tests may partially account for the complete absence of assessment that has been observed in many community
programs (Schroeder, 1995). However, increased pressure for accountability from the health care system and from state and federal funding agencies is prompting program planners to include assessment as a major part of their programs. The ability to provide relevant outcome measures, of course, depends upon the availability of suitable measurement instruments.

Summary

As life expectancy continues to increase, so too has the number of years people spend living with disabilities that compromise the quality of their lives. With the projected growth in the older adult population, preventing or delaying disability and promoting healthy lifestyles have become national goals. Research suggests that loss of function in later years, especially that associated with physical inactivity, can be reduced through accurate detection of physical impairment and proper adjustment in physical activity behavior. Historically, however, most test development to assess functioning in older adults has focused on the more frail elderly, primarily to identify their level of limitation and to determine the types of services needed. Few standardized tests have been developed that are suitable for assessing physical performance in the large population of independent, community-residing older adults. Especially lacking are tools that can measure performance on a continuum across the wide range of functioning in this segment of the population. Of particular concern is the ability to measure the underlying physiological parameters associated with common activities of daily living. Regardless of the cause (i.e., disease or disuse), abnormal declines in physical performance, if not remediated, can lead to disability.

Additional tests are needed to adequately assess older adults’ physical performance. Tests for older adults, as for younger populations, should be developed and evaluated based on well-established scientific standards. In addition to being socially acceptable to the participants and feasible for use in the intended setting, all assessment items should have acceptable test–retest reliability, should be appropriately valid, should have sufficient discrimination power to assess the population of interest, and should include standards of performance to facilitate interpretation of scores.

TEST DEVELOPMENT IN PROGRESS

In response to the need for additional tools to assess physical performance in older adults, we have developed a revised/recommended battery of tests, which is currently under review. The tests, which were designed to meet the objectives and guidelines discussed in this paper, include measures of upper body strength/endurance, lower body strength/endurance, aerobic endurance, shoulder flexibility, trunk/hamstring flexibility, and dynamic balance/agility. All tests are relatively quick and easy to administer, require minimal equipment and space, and can assess ability levels ranging from the borderline frail to the highly fit. Copies of the test battery can be obtained from either of the authors.1 In particular, those with access

---

1A copy of the test battery and instructions for participating in the collection of national norms can be obtained from Roberta Rikki or Jessie Jones, Division of Kinesiology and Health Promotion, California State University, Fullerton, Fullerton, CA 92834.
to older adult populations are invited to contact us about participating in the data collection for establishing national norms for the test items.

References


Acknowledgment
This work was supported by PacifiCare Health Systems.