Reproducibility of Physical Performance and Physiologic Assessments

FREDRIC D. WOLINSKY
University of Iowa, Iowa City VAMC
DOUGLAS K. MILLER
Indiana University, Regenstrief Institute for Health Care
ELENA M. ANDRESEN
University of Florida, Gainsville VAMC
THEODORE K. MALMSTROM
Saint Louis University
J. PHILIP MILLER
Washington University

Objective: We evaluate the test-retest stability of physical performance and physiologic assessments used in epidemiologic research. **Method:** Eighty subjects aged 50 to 65 were randomly selected from a probability sample of African Americans for test-retest assessments 5 to 45 days after baseline. Physical performance assessments included grip strength, chair stands, gait speed, and four standing-balance measures. Physiologic assessments included systolic and diastolic blood pressure, height, weight, body fat, and peak expiratory flow. **Results:** Intraclass correlations coefficients (ICCs) were .81 for grip strength, .72 for chair stands, .56 for gait speed, .60 for one-leg stand, .52 for semitandem stand, .58 for tandem stand with eyes closed, and .27 for tandem stand with eyes open. Except for blood pressure (ICCs of .51 and .55 for systolic and diastolic), the physiologic assessments had ICCs > .89. **Discussion:** Additional interviewer training may improve the reproducibility of the tandem stand with eyes open.

Keywords: *test-retest reliability; physical performance assessments; physiologic assessments; African Americans*

The importance of measuring functional status in both the clinical management and epidemiologic assessment of middle-aged and older

JOURNAL OF AGING AND HEALTH, Vol. 17 No. 2, April 2005 111-124
DOI: 10.1177/0898264304272784

adults is widely recognized (Williams, 1986). There is also general agreement that comprehensive assessments should cut across diseases and functional domains (Applegate, Blass, & Williams, 1990) and that these would complement clinical examinations (Tinetti & Ginter, 1988). As a result, scores of self-report and performance-based (Cress et al., 1995) functional status measures have been developed since the mid-1930s (Feinstein, Josephy, & Wells, 1986). What is not clear is whether self-report is preferable to performance, whether performance is preferable to self-report, or whether they are complementary (Myers, Holliday, Harvey, & Hutchinson, 1993; Reuben, Siu, & Kimpau, 1992; Rozzini, Frisoni, Bianchetti, Zanetti, & Trabucchi, 1993).

Despite the unresolved preference question, the use of physical performance assessments (e.g., grip strength, chair stands, gait speed, and balance tests) has been advocated for both cross-sectional and longitudinal research because they yield more objective results than self-reports (Guralnik, Branch, Cummings, & Curb, 1989). Physical performance assessments are now relatively common in major studies, including the Established Populations for the Epidemiologic Study of the Elderly (Guralnik et al., 2000); the Health, Aging, and Body Composition Study (Simonsick et al., 2001); the MacArthur Studies of Successful Aging (Seeman et al., 1994); the Women's Health and Aging Study (Ferrucci et al., 1997); and the Zutphen Elderly Study (Hoeymans, Feskens, van den Bos, & Kromhout, 1996).

The reproducibility (i.e., stability or test-retest reliability) of physical performance assessments, however, is not well established in community and minority populations. Most available studies are based on very small and nongeneralizable samples that have as few as 12 to 69 participants from convenience samples consisting primarily of Whites drawn from day care and other clinical settings (Daltroy et al., 1995; Fox, Felsenthal, Hebel, Zimmerman, & Magaziner, 1996; Gerety et al., 1994; Salen, Spangfort, Nygren, & Nordemar, 1994;

AUTHORS' NOTE: Supported by NIH Grant R01 AG-10436 to Douglas K. Miller. The opinions expressed here are those of the authors and do not necessarily reflect those of the NIH or any of the academic or governmental institutions involved. Address all correspondence to Fredric D. Wolinsky, the John W. Colloton Chair in Health Management and Policy, College of Public Health, The University of Iowa, 200 Hawkins Drive, E-205 GH, Iowa City, Iowa 52242; e-mail: fredric-wolinsky@uiowa.edu.

Thomas & Hageman, 2002; Winograd et al., 1994). Although the few larger studies provide more precise estimates, only a few physical performance measures have been considered in any particular study, and a wide range of reproducibility coefficients have been reported, with standing-balance tests generally faring the worst (Hoeymans, Wouters, Feskens, van den Bos, & Kromhout, 1997; Jette, Jette, Ng, Plotkin, & Bach, 1999; Shields, Enloe, Evans, Smith, & Steckel, 1995). Thus, further research on the reproducibility of physical performance assessments in representative, community-based samples, especially those that include or focus on minority participants, is warranted.

Similarly, some basic physiologic assessments (e.g., height and weight) of participants provide more objective information than self-reports and are easy to obtain. Other physiologic assessments (e.g., body fat, blood pressure, and peak expiratory flow), however, simply can not be obtained from self-reports. Making these physiologic assessments requires the use of more specialized equipment and more advanced interviewer training. As with physical performance assessments, the reproducibility of physiologic assessments in representative, community-based samples, especially those that include or focus on minority participants, warrants further study.

The purpose of this article is to examine the reproducibility of both physical performance and physiologic assessments. Between September 2000 and July 2001, 998 African Americans who were born in 1936 through 1950 were enrolled in the African American Health (AAH) project, a longitudinal study using a probability-based sample drawn from St. Louis, Missouri. The overall purpose of the AAH is to examine the disablement process and identify critical pathways at which intelligent interventional strategies can be targeted. The 3-year data collection cycle for the AAH involves a comprehensive in-home assessment in year 1 that includes numerous demographic, socioeconomic, psychosocial, and biomedical measures. This is followed by a brief telephone interview in years two and three for tracking and annual incidence monitoring. Then, the three-year cycle repeats itself. A random subsample of 114 of the 998 participants at baseline was selected for test-retest interviews. Both physical performance (i.e., grip strength; repeated chair stands; gait speed; and semitandem, tandem-with-eyes-open, tandem-with-eyes-closed, and one-leg

stands) and physiologic (i.e., blood pressure, height, weight, percentage body fat, and peak expiratory flow) assessments were included in the test-retest protocol.

Method

SAMPLE

The AAH has been described elsewhere (Miller et al., 2004; Wolinsky, Miller, Andresen, Malmstrom, & Miller, 2004). There were two sampling strata. The first involved a poor, inner-city area of St. Louis. The second area involved near northwest suburbs. In this stratum, census data were used to identify blocks having at least 10% African Americans. Sampling proportions were set to recruit approximately equal numbers of participants from both strata; but because the city stratum had fewer eligible participants, it was, therefore, oversampled relative to population size. Besides birth dates, inclusion criteria involved self-reported Black or African American race, Standardized Mini-Mental Status Examination scores ≥ 16 (Molloy, Silberfeld, & Darzins, 1996), and willingness to sign informed consent. All study design and procedural elements received approval from the Saint Louis University Institutional Review Board.

Participants received in-home, baseline evaluations that averaged 2.5 hr and were conducted by 50 interviewers. The overall response rate was 76% (77% in the inner-city stratum and 75% in the suburban stratum). Data for this article were taken from a substudy nested within the main design. In this substudy, 114 of the 998 participants who completed the baseline assessment were randomly selected for and asked to complete a test-retest interview within 5 to 45 days. Of these, 92 participants (81%) and 25 interviewers participated in the test-retest substudy, and, for 80 of these participants, the test and retest interviewers were matched (i.e., 12 participants had a different interviewer at test than at retest). There were no significant ($p < .05$) differences between the participants in the substudy versus those only in the main study in terms of age, sex, education, income, or any of the physical performance or physiologic assessments.

METHODOLOGICAL ASSUMPTIONS

Two assumptions must be met in order for test-retest reliability coefficients to reflect the stability of a measure (Cronbach, 1947). The first is that the construct being measured does not change during the measurement process. In terms of body composition (i.e., height, weight, and percentage body fat), strength (i.e., grip strength and chair stands), gait speed, and balance (i.e., one-leg, semitandem, tandem-with-eyes-open, and tandem-with-eyes-closed stands), the assumption of construct stability during the test-retest period is reasonable.

Stronger assumptions, however, must be made for blood pressure (both systolic and diastolic) and peak expiratory flow. That is, it is reasonable to expect greater intraindividual variation in blood pressure and peak expiratory flow because the underlying functions are subject to considerable influence from basic biologic variability as well as psychosocial and environmental factors beyond the control of the study. Although using the average from two automated sphygmomanometer assessments of blood pressure and the average from three peak flow assessments reduces this variation, it does not eliminate it. Thus, it is reasonable to expect that the reproducibility coefficients for blood pressure and peak flow measures may be lower than those for the other assessments.

The second assumption underlying the ability of test-retest reliability coefficients to reflect the stability of a measure is that the two observations are independent (Cronbach, 1947). At issue here is the length of the interval between the test and the retest (Nunnally, 1967). If too little time lapses between the test and the retest, the estimate of reproducibility may be biased upwards because participants may try to remember their prior answers (not relevant in this study) or because interviewers may try to remember their prior readings. Pragmatically, it is often assumed that the optimal test-retest interval is somewhere between several days and several weeks (DeVellis, 2003). Given the number of days in between the baseline (i.e., test) and the retest assessments in this study (mean = 18, median = 19, interquartile range = 13 to 22), the potential for participant and interviewer bias is marginal.

THE PHYSICAL PERFORMANCE ASSESSMENTS

Interviewers demonstrated each of the physical performance measures to the participants before the assessments were made. Grip strength was assessed as the average of three maximal effort trials using a hand-held dynamometer. After the interviewer demonstrated the technique, participants self-selected their stronger hand and performed the assessments while seated with the elbows of their chosen arms at 90-degree unobstructed angles and with the wrist in the neutral position between internal and external rotation. A straight back, sturdy chair without arms with the seat height appropriate for the participant's height was used for the chair stands measurement. Participants were instructed to cross their arms over their chests and complete five combinations of rises and returns as fast as they could. Electronic stopwatches accurate to one one-hundredth of a second were used. Start time was the interviewer's instruction to begin, and stop time was the peak of the fifth rise. Maximal time allowed for chair stands was 60 s. A 4-meter course was demarcated (54% of participants used a 3-meter course at baseline because of space constraints in their home) using a premeasured flat cord across which a starting line, a small end-of-course line (for the interviewer's benefit), and a stop line (one meter after the end-of-course line) were marked using nonmarring tape. Participants were instructed to walk at their normal pace, as if going to the store. The average time obtained from two trials was used, with a 50-s maximal time allowed. Although gait speed is usually reported in meters per second, it is appropriate in this reproducibility study to disregard course length inasmuch as all but three participants used the same length course at test and retest and the different course lengths for those three participants are off-setting (i.e., they introduce only a minimal amount of random error). Four balance assessments were used, with stopwatches recording timings, up to a maximum of 30 s. For the first three balance assessments, a line was made using nonmarring tape on the floor to facilitate participant alignment and interviewer observation. Using standard methods, participants then performed semitandem, tandem- with-eyes-open, and tandem-with-eyes-closed stands. The fourth balance assessment was a one-leg stand using the participant's self-selected leg.

THE PHYSIOLOGIC ASSESSMENTS

Two readings from an automated sphygmomanometer were used to measure systolic and diastolic blood pressure. In between these two assessments, interviewers loosened the cuff, recorded the first reading, asked participants six questions about their vision and hearing, and then retightened the cuff before taking the second reading. Height (without shoes) was assessed by having participants stand erect against a doorframe to which nonmarring tape had been applied. Interviewers used a plastic right angle to locate and mark the participant's height on the tape and then assessed the distance (in inches) from the floor to the tape mark using a retractable metal tape measure. Weight (without shoes and stockings) was measured by having the participant stand on a Tanita 2000 electric impedance scale that was placed on a flat, hard surface. Percentage body fat was adjusted internally by the Tanita 2000 for sex plus measured height and weight. Peak expiratory flow was assessed as the average of three trials using a standard flow meter with the participant standing erect. A 30-s rest period occurred before each trial, and the interviewer demonstrated the technique prior to participant assessment.

INTERVIEWER TRAINING

The Survey Research Center (SRC) at the University of Michigan collected the data for this study. After completing standard SRC interviewer-training modules, project staff received extensive study-specific training. This included an overview of the study purpose and design, rationales for all items selected, group and individual practice sessions in the computer-assisted survey components of the study, and extensive training and certification in the physical performance and physiologic assessment component. The latter was spread across sessions occurring across the 5-day study-specific training period to minimize fatigue and maximize retention and standardization. The performance- and physiologic-assessment-training protocol included the viewing (with discussion) of two videos in which two of the authors (Fredric D. Wolinsky and Douglas K. Miller) conducted the performance assessments properly and improperly, as well as repeated

group and individual hands-on demonstration and practice, culminating in a formal certification examination. Overall, the physical performance and physiologic assessment (including falls safety) training and certification process involved an average of 12 clock hours per interviewer.

ANALYSIS

Reproducibility was assessed using intraclass correlation coefficients (ICCs), as this is the method considered most appropriate (Deyo, Diehr, & Patrick, 1991). The preferred ICC for test-retest reliability is the 3,1 model (Shrout & Fleiss, 1979) when participants and interviewers are matched (i.e., the same interviewer conducts both the test and retest assessments on the same participant). As indicated above, 80 of the 92 test-retest assessments were matched, and the analysis reported here is limited to those matched assessments. The 3,1 ICCs were calculated using the two-way random effects model with a consistency definition for single measures (SPSS, 2004). Additional analyses (not shown) using all 92 test-retest assessments and the appropriate 2,1 ICC model that relaxes the matching constraint (two-way random effects with an absolute agreement definition) yielded nearly identical results. Intraclass correlation coefficients reflect the proportion of variance in the test and retest scores that are attributable to the true score (Shrout & Fleiss, 1979) and, thus, have a proportional-reduction-in-error interpretation (e.g., η^2 and r^2). Because Fleiss and Cohen (1973) have shown that the ICC is mathematically equivalent to the kappa and weighted kappa statistics, it is appropriate to interpret ICCs using established conventions for the kappa family of statistics. Kappa values in the .41 to .60 range are considered "moderate"; values in the .61 to .80 range are considered to be "substantial"; and values greater than .80 are considered to be "almost perfect" (Landis & Koch, 1977, p. 165).

Table 1
Baseline and Retest Means and Standard Deviations, and Intraclass Correlation Coefficients, for the Physical Performance Measures[a]

	Baseline		*Retest*			
Assessment	M	SD	M	SD	*ICC*	*Number of Cases*
Grip strength (kilogram)	32.8	12.5	33.2	13.6	.81	64
Chair stands (second)	13.1	5.2	12.8	5.6	.72	48
Gait speed (second)	4.7	1.2	4.8	1.3	.56	32
One-leg stand (second)	18.5	10.9	20.3	11.7	.60	53
Semitandem stand (second)	26.8	8.6	26.0	9.9	.52	63
Tandem stand with eyes open (second)	26.0	8.6	27.9	6.2	.27	54
Tandem stand with eyes closed (second)	14.2	11.7	15.4	12.3	.58	52

Note: ICC = intraclass correlation coefficients.
a. As indicated in the text, all of the timed physical performance tests were measured to the nearest one one-hundredth of a second by electronic stopwatches.

Results

DESCRIPTIVE

The mean age of the 80 matched participants in the test-retest substudy was 56.6 years old (range = 50 to 64); 39% were men; 39% were divorced or separated; 18% were widowed; and 13% were single. Average educational attainment was 12.4 years (range = 3 to 20). Thirty percent of participants reported having difficulty with one or more Activities of Daily Living (ADLs), 38% reported having difficulty with one or more Instrumental ADLs, 56% reported having one or more lower body limitations, and 30% reported having one or more upper body limitations. Thirty-one percent of participants had fallen at least once during the past year, and 17.5% had clinically relevant levels of depressive symptoms indicated by their 11-item Center for Epidemiological Studies—Depression scores (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993). Based on self-reports, 6.3% had cancer, 23.8% had diabetes, 33.8% had heart disease, and 8.8% had lung disease.

Table 2
Baseline and Retest Means and Standard Deviations, and Intraclass Correlation Coefficients, for the Basic Performance (and Physiologic) Assessment Measures

	Baseline		*Retest*			
Assessment	M	SD	M	SD	*ICC*	*Number of Cases*
Systolic blood pressure	141.4	25.5	132.6	25.0	.51	76
Diastolic blood pressure	83.1	12.9	81.2	10.8	.55	75
Height (in.)	66.3	3.3	66.4	3.3	.99	63
Weight (lb)	184.8	36.4	185.0	38.8	1.00	62
Percentage body fat	35.2	11.0	35.0	11.6	.93	60
Peak expiratory flow (l/min)	409.7	130.4	406.1	142.3	.89	43

Note: ICC = intraclass correlation coefficients.

REPRODUCIBILITY

Table 1 shows the means, standard deviations, and ICCs obtained for the physical performance measures. The ICC for grip strength is robust; the ICC for chair stands is substantial; and the ICCs for gait speed, the one-leg stand, the semitandem stand, and the tandem stand with eyes closed are moderate; but the ICC for the tandem stand with eyes open is unacceptable. Table 2 shows the same information for the physiologic measures. With the exception of systolic and diastolic blood pressure, all ICC values for the physiologic measures are > .80 and reflect near perfect agreement between the baseline and retest assessments. The ICCs obtained for blood pressure are moderate.

Discussion

With the exception of the tandem stand with eyes open, the reproducibility of the physical performance assessments was moderate to substantial (i.e., the ICCs ranged from .52 to .81). Thus, grip strength, repeated chair stands, gait speed, one-leg stands, semi-tandem stands, and tandem stands with eyes closed are appropriate for use in repeated measures analyses of longitudinal epidemiologic data. The ICC (.27) for tandem stands with eyes open, however, was clearly unacceptable. Therefore, post hoc analyses were conducted in an attempt to explain this apparent anomaly.

Given the average improvement (1.9 s) in the time participants were able to hold the tandem stand with eyes open, a familiarity (or learning) effect seemed plausible. If a familiarity effect was operative, however, the simple Pearson product—moment correlation coefficient between the test and retest timings should have been appreciably higher than the ICC shown in Table 1; it was not ($r = .28$). In fact, none of the Pearson product—moment correlation coefficients between the test and retest values differed by more than .025 from their respective ICC values for any of the physical performance or physiologic assessments.

The best plausible explanation that remains for the unacceptable reproducibility of the tandem stand with eyes open in the face of the moderate reproducibility of the tandem stand with eyes closed may reside in the differences between the average times that each could be held. As shown in Table 1, the average time that these same participants could hold the tandem stand with their eyes closed was only half of the time that they could hold the stand with their eyes open. Thus, the difficulty of holding the tandem stand with eyes closed may have constrained the opportunity for test-retest differences. Further investigation of this possibility is beyond the scope of these data.

Reproducibility for the physiologic assessments was very high (i.e., the ICCs ranged from .83 to 1.00), except for systolic and diastolic blood pressure, which were moderate (ICCs of .51 and .55, respectively). This was expected, inasmuch as intraindividual variation in blood pressure is well established and is known to be sensitive to biological variability and psychosocial and environmental factors beyond the control of the study. Indeed, the median intraindividual (i.e., test-retest) differences observed in this study for both systolic (8.8 mmHg) and diastolic (1.9 mmHg) blood pressure are quite comparable to those reported from repeated assessments conducted on the same day (Armitage, Fox, Rose, & Tinker, 1966). Nonetheless, the results do raise some concerns about the appropriateness of blood pressure assessments for repeated measures analyses in longitudinal studies.

In concluding this article, it is worth noting that some studies categorize physical performance assessment timings. The reproducibility analyses reported here were not conducted on categorized timings. There are several reasons for this. First, our participants are middle-aged African Americans. The categorized timings reported in other

applications are based on older samples of mostly White participants. Although we could have used comparable categorization strategies (i.e., five categories based on inability to perform the assessment, plus the four quartiles of performed timings), such an approach is ill advised in longitudinal research involving repeated measures analyses. Second, even if appropriate categorized timing norms were available, conducting the reproducibility analyses on them would create differential opportunities for change among participants depending on how close their test and re-test timings were to the categorical cut points. Finally, it is ultimately the reproducibility of the timings that is clinically important, not their subsequent and arbitrary transformation.

REFERENCES

Applegate, W. B., Blass, J. P., & Williams, T. F. (1990). Instruments for the functional assessment of older adults. *New England Journal of Medicine*, *322*, 1207-1214.

Armitage, P., Fox, W., Rose, G. A., & Tinker, C. M. (1966). The variability of measurement of casual blood pressure: II. Survey experience. *Clinical Sciences*, *30*, 337-344.

Cress, M. E., Schechtman, K. B., Mulrow, C. D., Fiatarone, M. A., Gerety, M. B., & Buchner, D. M. (1995). Relationship between physical performance and self-perceived physical function. *Journal of the American Geriatrics Society*, *43*, 93-101.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*, 1-16.

Daltroy, L. H., Phillips, C. B., Eaton, H. M., Larson, M. G., Partridge, A. J., Logigian, M., et al. (1995). Objectively measuring physical ability in elderly persons: The physical capacity evaluation. *American Journal of Public Health*, *85*, 558-560.

DeVellis, R. F. (2003). Scale development: Theory and applications (2nd ed.). Newbury Park, CA: Sage.

Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures. *Controlled Clinical Trials*, *12*, 142S-158S.

Feinstein, A. R., Josephy, B. R., & Wells, C. K. (1986). Scientific and clinical problems in indexes of functional disability. *Annals of Internal Medicine*, *105*, 413-420.

Ferrucci, L., Guralnik, J. M., Buchner, D., Pahor, M., Harris, T., Corti, M. C., et al. (1997). Departures from linearity in the relationship between measures of muscular strength and physical performance of the lower extremes: The Women's Health and Aging Study. *Journal of Gerontology: Medical Sciences*, *52A*, M275-M285.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613-619.

Fox, K. M., Felsenthal, G., Hebel, J. R., Zimmerman, S. I., & Magaziner, J. (1996). A portable neuromuscular function assessment for studying recovery from hip fracture. *Archives of Physical Medicine and Rehabilitation*, *77*, 171-176.

Gerety, M. B., Mulrow, C. D., Tuley, M. R., Hazuda, H. P., Lichtenstein, M. J., Bohannon, R., et al. (1994). Development and validation of a physical performance instrument for the functionally impaired elderly: The physical disability index (PDI). *Journal of Gerontology: Medical Sciences*, *48*, M33-M38.

Guralnik, J. M., Branch, L. G., Cummings, S. M., & Curb, J. D. (1989). Physical performance measures in aging research. *Journal of Gerontology: Medical Sciences*, *44*, M141-M146.

Guralnik, J. M., Ferrucci, L., Pieper, C. F., Leveille, S. G., Markides, K. S., Ostir, G. V., et al. (2000). Lower extremity function and subsequent disability: Consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. *Journal of Gerontology: Medical Sciences*, *55A*, M221-M231.

Hoeymans, N., Feskens, E. J. M., van den Bos, G. A. M., & Kromhout, D. (1996). Measuring functional status: Cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990-1993). *Journal of Clinical Epidemiology*, *49*, 1103-1110.

Hoeymans, N., Wouters, E. R. C. M., Feskens, E. J. M., van den Bos, C. A. M., & Kromhout, D. (1997). Reproducibility of performance-based and self-reported measures of functional status. *Journal of Gerontology: Medical Sciences*, *52A*, M363-M368.

Jette, A. M., Jette, D. U., Ng, J., Plotkin, D. J., & Bach, M. A. (1999). Are performance-based measures sufficiently reliable for use in multicenter trials? *Journal of Gerontology: Medical Sciences*, *54A*, M3-M6.

Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the CES-D depression symptoms index. *Journal of Aging Health*, *5*, 179-193.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Miller, D. K., Malmstrom, T. K., Joshi, S., Andresen, E. M., Morley, J. E., & Wolinsky, F. D. (2004). Clinically relevant levels of depressive symptoms in community-dwelling middle-aged African Americans. *Journal of the American Geriatrics Society*, *52*, 741-748.

Molloy, D. W., Silberfeld, M., & Darzins, P. (1996). Measuring capacity to complete an advance directive. *Journal of the American Geriatrics Society*, *44*, 660-664.

Myers, A. M., Holliday, P. J., Harvey, K. A., & Hutchinson, K. S. (1993). Functional performance measures: Are they superior to self-assessments? *Journal of Gerontology: Medical Sciences*, *48*, M196-M206.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Reuben, D. B., Siu, A. L., & Kimpau, S. (1992). The predictive validity of self-report and performance-based measures of function and health. *Journal of Gerontology: Medical Sciences*, *47*, M106-M110.

Rozzini, R., Frisoni, E. B., Bianchetti, A., Zanetti, O., & Trabucchi, M. (1993). Physical performance test and Activities of Daily Living scales in the assessment of health status in elderly people. *Journal of the American Geriatrics Society*, *41*, 1109-1113.

Salen, B. A., Spangfort, E. V., Nygren, A. L., & Nordemar, R. (1994). The disability rating index: An instrument for the assessment of disability in clinical settings. *Journal of Clinical Epidemiology*, *47*, 1423-1434.

Seeman, T. E., Charpentier, P. A., Berkman, L. F., Tinetti, M. E., Guralnik, J. M., Albert, M., et al. (1994). Predicting changes in physical performance in a high-functioning elderly cohort: MacArthur Studies in Successful Aging. *Journal of Gerontology: Medical Sciences*, *49*, M97-M108.

Shields, R. K., Enloe, L. J., Evans, R. E., Smith, K. B., & Steckel, S. D. (1995). Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. *Physical Therapy*, *75*, 169-179.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Simonsick, E. M., Newman, A. B., Nevitt, M. C., Kritchevsky, S. B., Ferrucci, L., Guralnik, J. M., et al. (2001). Measuring higher level physical function in well-functioning older adults: Expanding familiar approaches in the Health ABC study. *Journal of Gerontology: Medical Sciences*, *56A*, M644-M649.

SPSS. (2004). SPSS 12 brief guide. Chicago: Author.

Thomas, V. S., & Hageman, P. A. (2002). A preliminary study on the reliability of physical performance measures in older day care center clients with dementia. *International Psychogeriatrics*, *14*, 17-23.

Tinetti, M. E., & Ginter, S. F. (1988). Identifying mobility dysfunctions in elderly patients: Standard neuromuscular examination or direct assessment? *Journal of the American Medical Association*, *259*, 1190-1193.

Williams, M. E. (1986). Geriatric assessment. *Annals of Internal Medicine*, *104*, 720-721.

Winograd, C. H., Lemsky, C. M., Nevitt, M. C., Nordstom, T. M., Stewart, A. L., Miller, D. J., et al. (1994). Development of a physical performance and mobility examination. *Journal of the American Geriatric Society*, *42*, 743-749.

Wolinsky, F. D., Miller, D. K., Andresen, E. M., Malmstrom, T. K., & Miller, J. P. (2004). Health-related quality of life among middle-aged African Americans. *Journal of Gerontology: Social Sciences*, *59B*, S118-S123.