Journal of Gerontology: MEDICAL SCIENCES
1998, Vol. 53A, No. 4, M295–M300

Reliability of Physical Performance and Self-Reported Functional Measures in an Older Population

Ira B. Tager, Ann Swanson, and William A. Satariano

Division of Public Health Biology and Epidemiology, School of Public Health, University of California at Berkeley.

Background. Functional assessments and direct measures of physical performance are standard components of community-based studies of older populations. Estimates of the reliability of these measures are necessary for the assessment of functional change.

Methods. The reproducibility of 13 measures of self-reported function and 11 direct measures of physical performance was assessed. A sample of subjects (N = 199; ≥ 55 yrs) was selected from a larger population-based cohort. Subjects were tested in their homes twice, 48 hours apart, by the same interviewer to replicate study conditions. Age-adjusted kappa statistics were used to assess the reliability of measures of physical function; product moment correlation (Pearson r) and intraclass correlation coefficients (ICC) were used to assess direct measures of performance. A repeated measures model was used to assess learning or practice effects of performance, adjusted for age, sex, general health, and cognitive function.

Results. Age-adjusted kappa statistics were ≥ .60 for most self-reported items. ICC ranged from .63 to .92. Significant improvements (practice effects) were found for the chair stand, walking speed, and the 360° turn. Measures of grip strength, reaching down, and hand dexterity were found to be reliable, with no significant test effect.

Conclusion. Three commonly used measures (chair stand, walking speed, and 360° turn) may be less reliable than previously reported. Sample sizes that may be needed to detect change in these areas of performance may be larger than previously estimated given this level of imprecision. Future studies of reproducibility should assess both the level of agreement and the presence of possible practice effects.

FUNCTIONING, defined generally as the ability to complete physical and cognitive tasks necessary for independence and adaptation to the environment, is central to an understanding of the health of older populations (1–3). Although functioning tends to decline with age, there is considerable variability within age groups (4–7). This variability depends on a number of factors, including age, sex, the prevalence of chronic health conditions, past and current health practices (4–7), and errors in reporting and measurement.

Quantitative estimates of the reliability (reproducibility) of measures of function are necessary for the assessment of functional changes over time. Large random errors in tests of function, on average, will lead either to underestimation of the magnitude of changes or to the complete obscuring of the occurrence of changes in function with age and disability or to improvements with targeted interventions (8). The reliability both of self-reported physical function and direct measures of physical performance has been examined in two large epidemiologic studies that used a 2- to 3-week interval between test and retest (9,10). Pearson correlations ranged from .58–.73 for a modified activities of daily living (ADL) index, a Rosow-Breslau index, and a Nagi index (9). For direct measures of physical performance (balance, chair stand, time for a fast walk, writing a signature), Pearson correlations between .61 and .91 have been reported (10). However, the use of a 2- to 3-week interval limits the interpretation of these results, as "real" acute changes in the status of these elderly subjects could have occurred, and this possibility is not specifically addressed in these studies.

As part of a longitudinal, population-based study of physical functioning in elders, a study was undertaken to obtain estimates of the reliability of measures of self-reported physical function and direct measures of physical performance. To minimize the possibility of confounding of the estimates by "real" acute changes in status, the test-retest interval was restricted to 48 hours (11).

Methods

Subjects and Recruitment

The Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a community-based longitudinal study of age-related changes in physical functioning, physical activity, and fitness in persons ≥ 55 years who live in the city of Sonoma, California, and its environs. A community-based census identified 3,057 age-eligible individuals; 2,092 (68.4%; 1,246 females, 846 males) agreed to participate in the study and were enrolled between May 1993 and December 1994. The age distribution of these subjects was similar to that of the 1990 census data for persons ≥ 55 years who resided in the study community. Protocols were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley, and the Human Subjects Committee at the University of California, San Francisco.

Between the last week of February and the end of April 1995, 200 subjects were selected for inclusion in a substudy to evaluate the reproducibility of measures of self-reported function and direct physical performance. To assure adequate precision in "younger" and "older" subjects, subjects ≥ 70 years were to be oversampled (target n = 120) relative to subjects between the ages 55–69 years (target n = 80). Similarly, the target number of males (100) represented an oversampling relative to the percentage in the full study (40.4%).

Subjects were recruited during a scheduled telephone interview for an interval evaluation of morbidity and mortality. Subjects were eligible to participate if they were ambulatory without assistance (no wheelchair, cane, or walker), had not had any cataract surgery in the previous 6 weeks, had not had hand or arm injury or surgery in the previous 3 months, and did not have an acute medical condition that would prevent them from attempting the direct assessments of physical performance. Subjects were recruited consecutively until the quota for each of the four age-sex categories was filled.

All subjects were tested 48 hours apart; 98% of the retests occurred within one hour of the time of day of the first test, and all occurred within two hours of the time of day of the first test (range of total time elapsed between first and second test = 47–49 hours). Testing was conducted in the subjects' homes to replicate the testing conditions of the actual study. Each participant was tested by the same interviewer at each test session; four interviewers carried out the testing (Interviewers 1–4: 50.8%, 15.1%, 18.6%, 15.6%, respectively).

Physical Function (Self-Reported)

Measures of physical function, described by Rosow and Breslau (12) and Nagi (13), were based on self-reported assessments of level of difficulty in the performance of tasks that reflect lower- and upper-body strength, balance, and fine dexterity. Additional items that reflect more detailed assessments of lower-body function (getting up from a stooped position and getting up from a seated position) were developed for the SPPARCS project (see Table 1).

Each of the 13 self-reported physical functional measures has nine possible responses, four of which reflect level of difficulty for people who perform the task ("no difficulty," "a little difficulty," "some difficulty," "a lot of difficulty"), two responses for people who do not perform the task ("don't do on doctor's orders," "don't do because unable to do"), two final responses for people who "never" do the activity, and a response for people who "don't know." These latter two categories were excluded and the categories for nonperformance were grouped. The four responses that reflect levels of difficulty for people performing the task were maintained.

Analysis of self-reported measures.—The age and sex-specific percentages of exact agreement (same response category) and near agreement (± 1 category difference in response) were calculated based on the five response categories for each of the 13 physical functional measures.

Categorical measures of physical function were evaluated with the kappa statistic (14). Most respondents indicated that they did not have difficulty performing the task. Because a lower prevalence of responses can result in symmetrical imbalances of the marginal totals and high expected agreement, both of which can "falsely" lower kappa, the response categories were classified into two categories: "no difficulty" versus "difficulty," i.e., any report of difficulty (15). Separate kappa statistics were calculated for the two age groups within each sex and evaluated for homogeneity across age strata.

Direct Measures of Physical Performance

The measures included in this study represent direct assessments that have been used as part of the standard protocols for community-based epidemiological studies of older populations (Table 1). All measures, except grip strength and the 360° turn, were based on the timed performance of specific tasks.

Analysis of direct measures.—All measures were continuous variables, except for standing balance and the 360° turn, two ordinal-level measures. A weighted kappa statistic was used to examine standing balance (14). Weights were the default weights of the SAS procedure that gave results

Table 1. Measures of Physical Function and Physical Performance

Physical Function (Self-Report)
Nagi items (13)
Pushing heavy objects
Lifting < 10 lbs.
Lifting > 10 lbs.
Stooping, crouching, kneeling
Reaching or extending right arm
Reaching or extending left arm
Writing or handling small objects
Sitting for 1 hour or more
Standing in place for 15 or more minutes
Rosow-Breslau items (12)
Walking up and down stairs
Walking 3 or more city blocks
SPPARCS items
Getting up from a stooped position
Standing up from a seated position
Physical Performance (Direct Assessment)
Chair stand*
Side-by-side stand*
Semi-tandem stand*
Tandem stand*
One-legged stand*
10-foot walk*
360° turn*
Hand dexterity†
Reaching down (3 replications, each separated by several minutes)*
Dominant hand grip strength (3 replications, each separated by several minutes)*
Nondominant hand grip strength (3 replications, each separated by several minutes)*

*References 10, 11, 25, 26.

†Personal communication, M. Nevitt, July 1992. Time required to turn 5 coins (quarters) in sequence.

virtually identical to those with log-linear models that were used in preliminary analyses (16).

For reaching down and grip strength, three replicates were available at each test session; for all other direct assessments, a single measurement was obtained at each visit. To compare the present data with data from previous studies (10), Pearson correlations were obtained. However, standard correlation coefficients are not optimal measures of reliability (17,18). Therefore, intraclass correlation coefficients (ICC) [within-subject variance/(within-subject + between-subject variance)] were obtained from variance components derived from a nested one-way random effects ANOVA (grip strength and reaching down) (Proc Nested of SAS) (19); and a linear model (Proc Mixed of SAS) (20) with "subjects" as a repeated measure that adjusted for the effects of age and sex.

The differences in test results between the first and second visits were used to evaluate possible learning or practice effects. Plots of test differences versus the means of both tests were evaluated (17). Test-retest differences were adjusted for age, sex, general health [report of a history of one or more of: cardiovascular disease, cerebrovascular disease, chronic obstructive pulmonary disease/asthma, cancer (excluding non-melanoma skin cancer), diabetes mellitus, liver or renal disease], and cognitive function [modified Mini-Mental State Examination; MMSE (21)] in a repeated measures model with visit as a main effect (Proc Mixed of SAS) (20). Preliminary analysis revealed that virtually all subjects were oriented to time and place. To enhance the utility of the MMSE for use in a noninstitutionalized population, a subset of six items that were found to provide a reasonable distribution of cognitive function in the sample was used. The items included questions and tasks in which 10% or more of the subjects responded or performed incorrectly on either visit. The values were grouped into two categories: lowest quartile (scores 0–14) and upper three quartiles (scores 15–18).

Results

Study Sample

One-hundred ninety-nine subjects completed both interviews. Forty-two percent (n = 83: 41 females, 42 males) of the participants were between ages 55–69 years and 60% (n = 116: 58 females, 58 males) were ≥ 70 years. No statistically significant age difference (< 70 and ≥ 70 years) was found by participant status in the reproducibility study for either males or females. Therefore, age adjustment was not employed for comparisons between these groups. There was no significant difference in general health status between participants and nonparticipants for either men or women.

Of the 24 measures of physical function and physical performance, differences were found between participants and nonparticipants ($p \leq .10$) in five of the measures for females and in eight of the measures for males (Table 2). Most of the differences were found in the self-reported measures of physical function (5/13 for females and 6/13 for males); the pattern of difference was not consistent between sexes. Participants were more likely than nonparticipants to report difficulty with lower-body function (stooping, getting up from a stooped position, standing, and sitting); nonparticipants were more likely than participants to report difficulty with upper-body function (lifting and reaching).

Physical Function

Percentages of exact and near agreement for each self-reported measure of physical function were assessed by age and sex (Table 3). The percentage of females < 70 years who gave the same response at Visit 1 and Visit 2 (exact agreement) ranged from a high of 95.1% for "reaching the left arm" and "sitting for one hour or more" to 68.3% and 63.4% for "stooping" and "getting up from a stooped position," respectively. For females aged ≥ 70, the percentage of exact agreement ranged from 91.4% for reaching or extending either arm to 67.2% for "getting up from a stooped position." Among males < 70 years, reports of difficulty with "pushing objects" had the highest level of exact agreement (97.6%) and difficulty "stooping," the lowest (78.6%). In contrast, for males ≥ 70, exact agreement ranged from a high of 96.6% for "reaching or extending the left arm" to a low of 70% for "stooping." Most of the percentages of near agreement (± 1 response category) were high; in only three cases was the percentage under 90% [females < 70: stooping (82.9%); females ≥ 70: stooping (89.7%), getting up from a stooped position (86.2%)]. This high level of agreement was due primarily to the large percentage of subjects who reported "no difficulty" at both visits. For 9 of the 13 items, 73% or more of the subjects reported "no difficulty" at both visits. Lower percentages were found for the two stooping items (38%; 52%) and for the two standing items (61%; 69%).

Kappa statistics for self-reported measures of function ranged from a low of .38 among males for reported difficulty

Table 2. Comparison of Function Measures Between Subjects for Reproducibility Study and the Remaining Study Participants; Percent Reporting Difficulty*

	Females Reproducibility Status		Males Reproducibility Status	
Function Measures	Yes (n = 99)	No (n = 1146)	Yes (n = 100)	No (n = 746)
Stooping	50.5	47.5	47.0†	38.3
Getting up from a stooped position	69.7	65.8	60.0†	48.1
Lifting < 10 lbs	9.1†	17.8	3.0†	7.5
Extending dominant hand	5.1†	11.6	2.0†	9.4
Extending nondominant hand	3.0†	9.2	4.0†	9.7
Sitting for 1 hour or more	17.2	22.9	22.0†	14.2
Standing up from a chair	43.4†	33.8	31.0	26.0
Walking up or down stairs	22.2†	31.2	21.0	21.9
Dominant hand grip strength (median kg)	(n = 97) 22.7	(n = 1110) 23.3	(n = 100) 43.3‡	(n = 725) 40.7

*Only measures with $p \leq .10$ for at least 1 comparison are reported.
†χ^2.
‡Kruskal Wallis ANOVA (χ^2 approximation).

Table 3. Test-Retest Agreement for Self-Report Measures of Physical Function

	Percentage Exact Agreement				Percentage Near Agreement*			
	Female†		Male†		Female		Male	
Function Measures	55–69	70+	55–69	70+	55–69	70+	55–69	70+
Nagi items								
Pushing objects	82.1	76.8	97.6	83.9	92.3	94.7	97.6	94.6
Lifting < 10 lbs	95.0	91.2	92.9	91.4	100.0	94.7	100.0	96.6
Lifting > 10 lbs	85.4	77.6	88.1	89.7	98.0	98.0	100.0	94.8
Stooping	68.3	70.7	78.6	70.7	82.9	89.7	95.2	96.6
Reaching right arm	92.7	91.4	92.9	94.8	100.0	94.8	97.6	100.0
Reaching left arm	95.1	91.4	92.9	96.6	100.0	96.6	100.0	96.3
Writing or handling small objects	82.9	81:0	92.9	91.4	97.6	98.3	97.6	98.3
Sitting for 1 hour or more	95.1	85.7	88.1	75.9	100.0	91.1	95.2	98.3
Standing in place for 15 or more minutes	82.9	73.5	94.6	82.4	94.3	91.8	94.6	96.1
Rosow-Breslau items								
Walking up or down stairs	87.8	85.2	90.5	83.9	97.6	94.5	97.6	98.2
Walking 3 or more city blocks	85.0	90.7	95.2	91.4	95.0	98.2	100.0	96.6
SPPARCS items								
Getting up from a stooped position	63.4	67.2	80.5	69.0	90.2	86.2	92.7	96.6
Standing up from a seated position	85.4	74.1	83.3	75.9	97.6	93.1	92.6	94.8

*± 1 difference in response categories (see Methods).
†Ages 55–69: 41 females, 42 males; Ages 70+: 58 females, 58 males.

in lifting items under 10 pounds to a high of .92 among females for reported difficulty in reaching or extending the left arm (Table 4). There was no consistent pattern of reproducibility by sex across the individual items. The kappa statistic was ≥ .60 for 12 of 13 measures for females and 11 of 13 for males.

Direct Measures of Physical Performance

The age-adjusted, weighted kappa for the measure of standing balance was .57 (95% CI: .40, .75) for females and .47 (95% CI: .27, .67) for males.

The test-retest correlations for continuous measures ranged from .67 for reaching down to .96 for dominant hand grip strength (Table 5). Only for reaching down, chair stand, and hand dexterity were the Pearson r values of similar magnitude to the ICC values. For walking speed and hand grip strength, the ICCs were lower than the Pearson correlations; for the 360° turn, the ICC was larger than the Spearman correlation.

There was a statistically significant improvement ($p \leq .05$) in scores for three of the six measures between Visits 1 and 2 (the chair stand, walking speed, and the 360° degree turn; Table 5). Differences in scores between Visits 1 and 2 did not vary significantly by age, sex, general health, or cognitive function. No test effect was evident for grip strength, reaching down, and the hand dexterity test. Figure 1 presents the relationship between the test difference (Visit 2 – Visit 1) and the mean of the results for both visits for tests of grip strength (mean difference not different from zero) and walking speed (mean difference .08 feet/second). For both measures, there is wide scatter of the differences around a mean difference of zero, and there is little evidence that the differences are systematically related to the means (product-moment correlations between difference and means: dominant handgrip = .05; walking speed = .14).

Table 4. Age-Adjusted Kappa Statistics (95% CI) for Self-Reported Physical Function*

Functional Measure	Females	Males
Nagi items		
Pushing heavy objects	.75 (.61–.89)	.81 (.66–.96)
Lifting < 10 lbs	.60 (.32–.87)	.38 (.04–.72)
Lifting > 10 lbs	.72 (.57–.87)	.71 (.53–.90)
Stooping	.58 (.42–.74)	.65 (.50–.81)
Reaching right arm	.88 (.75–1.02)	.69 (.41–.98)
Reaching left arm	.92 (.77–1.07)	.85 (.64–1.06)
Writing or handling small objects	.76 (.62–.90)	.79 (.60–.98)
Sitting for 1 hour or more	.80 (.66–.95)	.47 (.24–.69)
Standing in place for 15 or more minutes	.62 (.45–.78)	.77 (.64–.91)
Rosow-Breslau items		
Walking up or down stairs	.69 (.52–.85)	.78 (.63–.93)
Walking 3 or more city blocks	.75 (.59–.91)	.95 (.84–1.05)
SPPARCS items		
Getting up from a stooped position	.60 (.44–.75)	.78 (.66–.90)
Standing up from a seated position	.67 (.51–.83)	.64 (.48–.80)

*Each measure expressed as dichotomy: "difficulty" vs "no difficulty." See text for explanation.

Discussion

This investigation presents an in-depth evaluation of the reliability (reproducibility) and agreement for self-reported and direct measures of physical performance based on a 48-hour test-retest protocol. The results indicate that a number of the measures are subject to considerable random error and/or lack of agreement.

The apparent high level of exact agreement that was observed for many measures of self-reported physical functioning in this study (Table 3) and other studies (9) was

Table 5. Reproducibility of Physical Performance Measures

Test	Correlation		(Visit 2–Visit 1) Mean Difference (± *SD*)
	Pearson	Intraclass	
Dominant hand grip strength (kg)	.96	.79	.15 (± 3.5)
Reaching down (sec)	.67	.63	−.05 (± .72)
Chair stand (sec)	.81	.82	−.83 (± .15)*
Hand dexterity (sec)	.80	.79	−.22 (± .13)
360° turn (number of steps)†	.82	.92	−.14 (± .06)*
Walking speed (ft/sec)	.93	.78	.08 (± .02)*

*$p < .05$.
†Spearman correlation.

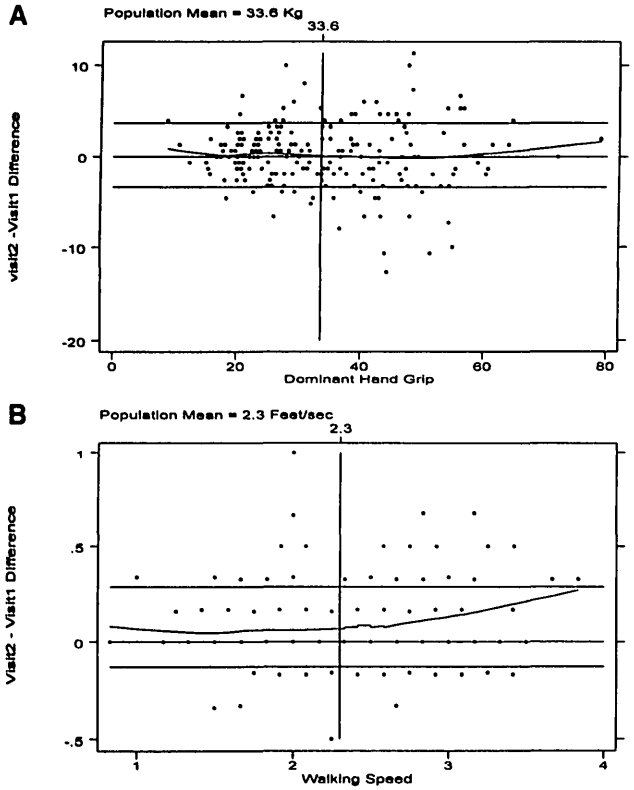

Figure 1. Relationship between (Visit 2 – Visit 1) test differences and the mean of the (Visit 2 + Visit 1) test results for grip strength (**A**) and walking speed (**B**). Vertical line is the mean test value for the study population. Horizontal lines above and below the "0" line represent ± 1 *SD* around the observed mean difference. Curved line is the lowess smooth estimate of the overall mean for the individual data points.

largely the result of the fact that most subjects reported no difficulty for the particular task. Except for questions that related to stooping (38%, 51%, "no difficulty"), 61–89% of the responses to other items were in the "no difficulty" category; and the percentage of subjects who reported no difficulties corresponds directly with the percentage of exact agreement (data not shown). Smith and colleagues (9) reported high levels of agreement for Katz (ADL) items in the East Boston component of Established Populations for Epidemiologic Studies of the Elderly (EPESE) but do not indicate the percentage of subjects who reported no problems with the given tasks. However, in the baseline survey of the East Boston study, the percentage of subjects who reported that they did not require assistance in ADL tasks ranged from 87.3% for bathing to 94.8% for using the toilet (22). The present study (Table 3) and the Smith study (9, Table 2) observed lower levels of agreement (similar magnitude between studies) for items derived from the Rosow-Breslau and Nagi items. In the present study, when responses were treated as dichotomies ("difficulty," "no difficulty") to balance the marginal distributions (15), a lower level of agreement between tests was observed, with the kappa statistics ≤ .80 for 11/13 measures in females and 10/13 for males (Table 4).

Very little data are available on the reproducibility of direct measures of physical performance that are expressed as continuous variables (10), and these data are given as product-moment correlation coefficients. Product-moment correlations are not measures of agreement per se (17,18) and mix estimates of between- and within-subject variability. Moreover, such correlations are influenced by the range of values. Use of the ICC more directly evaluates within-subject variability in a variance components framework (16). Age-adjusted within-subject variability represented ≥ 20% of the test-retest variability [ICC ≤ .80 for all direct measures except chair stand and the number of steps to complete a 360° turn (Table 5)]. Within-subject variability was greater in the older age group, but sex generally did not affect the variability, with the exception of significantly reduced grip strength in women.

Evidence of practice effects was clearly evident for three of the six direct measures (Table 5). Although the overall changes were small, a considerable number of subjects showed relatively large differences between the first and second tests, even for tests whose mean difference was not significantly different from zero (Figure 1A). The importance of practice effects has not been given sufficient attention in studies published to date. Unfortunately, with time intervals of 2 weeks or more between assessments, it has not been possible in previous studies to distinguish between practice effects and actual changes in performance that are related to changes in functional status.

Functional assessments and direct measures of physical performance have become standard components of study protocols in community-based studies of older populations (1–3). Together, these measures have been used to characterize the health and functional status of older populations and, most recently, have been used to identify states of "preclinical disability" (reduced levels of physical performance that have not as yet resulted in reports of difficulty or a loss of independence) (23). In general, direct measures have been used to identify different levels of performance among people who report no difficulty in areas of function (24). Direct measures have been further recommended as being able to provide more precise assessments of change in functional performance in aging populations than can be provided by self-report alone (25). Although direct measures hold the promise of providing more detailed and complete assessments of functioning and performance, the data from the present investigation indicate that three commonly

used measures may be less reliable than previously supposed (26). The sample sizes that may be needed by such studies to detect change over time may have been underestimated (8), given the apparent underestimation of the imprecision of these measures. Existing studies may be biased toward the null (27) or in unknown directions when multiple factors are measured with error (28). The problem is further compounded when the data generated from these measures are grouped into categories or summarized as function or performance scores (10,24). Indeed, two of the measures that showed the strongest test effect (chair stand and walking speed) are included as components in a general measure of lower-body strength (24).

Conclusion

In summary, several commonly used measures of performance and function may not be reliable as originally reported, a problem compounded by the generation of summary scales based on these measures. The development of more refined and integrated measures of function and performance depends in large part on the completion of comprehensive assessments of reliability. More detailed studies are needed to examine both the level of overall agreement, as measured by the ICC, and the presence of possible test effects. Future research also should evaluate the utility and feasibility of completing multiple replicates of tests at the same visit to provide more stable estimates of performance, as was done in the current study for grip strength and reaching down. In general, replicate assessments should be considered for those measures in which fatigue and time will not adversely affect the subject's performance. Finally, as new measures of physical performance and function are developed, careful evaluation of the reproducibility of these measures must be conducted.

ACKNOWLEDGMENTS

This work was supported by grant R01 AG-09389 from the National Institute on Aging.

The authors express their appreciation to Jackie Smith, Patti LeBlanc, Jan Hansen, Linda Barr, Janet Thiessen, Cynthia Fleischer, Joan Stier, and Virginia Weisel, who recruited subjects and conducted interviews; to Long Ngo for statistical consultation; to Bao Nguyen for data management; and to Katya Gomez for secretarial assistance. The authors also thank the subjects and the physicians of Sonoma, California, and the Sonoma Valley Hospital, without whose cooperation this research would not have been possible.

Address correspondence to Dr. Ira B. Tager, Division of Public Health Biology and Epidemiology, School of Public Health, 140 Earl Warren Hall, University of California at Berkeley, Berkeley, CA 94720-7360.

REFERENCES

1. Rowe JW. Health care of the elderly. *N Engl J Med.* 1985;312:827–835.
2. Williams TF. Comprehensive functional assessment: an overview. *J Am Geriatr Soc.* 1983;31:637–641.
3. Applegate WB, Blass JP, Williams TF. Instruments for the functional assessment of older patients. *N Engl J Med.* 1990;322:1207–1214.
4. Mor V, Murphy J, Masterson-Allen S, et al. Risk of functional decline among well elders. *J Clin Epidemiol.* 1989;42:895–904.
5. Guralnik JM, Kaplan GA. Predictors of healthy aging: prospective evidence from the Alameda County Study. *Am J Public Health.* 1989;79:703–708.
6. Harris T, Kovar MG, Suzman R, Kleinman JC, Feldman JJ. Longitudinal study of physical ability in the oldest old. *Am J Public Health.* 1989;79:698–702.
7. Guralnik JM, LaCroix AZ, Abbot, RD, et al. Maintaining mobility in late life: 1. Demographic characteristics and chronic conditions. *Am J Epidemiol.* 1993;137:845–857.
8. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Oxford, UK: Clarendon Press, 1994.
9. Smith LA, Branch LG, Scherr PA, et al. Short-term variability of measures of physical function in older people. *J Am Geriatr Soc.* 1990;38:993–998.
10. Seeman TE, Charpentier PA, Berkman LF, et al. Predicting changes in physical performance in a high-functioning elderly cohort: MacArthur Studies of Successful Aging. *J Gerontol Med Sci.* 1994;49:M97–M108.
11. Rossiter-Fornoff JE, Wolf SL, Wolfson LT, Buchner DM, the FICSIT Group. A cross-sectional validation study of the FICSIT common data base static balance measures. *J Gerontol Med Sci.* 1995;50A:M291–M297.
12. Rosow I, Breslau N. A Guttman health scale for the aged. *J Gerontol.* 1966;21:556–559.
13. Nagi SZ. An epidemiology of disability among adults in the United States. *Milbank Q.* 1976;54:439–468.
14. Fleiss JL. *Statistical Methods for Rates and Proportions.* New York: John Wiley, 1981.
15. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol.* 1990;543–549.
16. Agresti A. Modeling for ordinal scales. *Biometrics.* 1988;44:539–548.
17. Bland JM, Altman DE. Comparing two methods of clinical measurement: a personal history. *Int J Epidemiol.* 1995;24 (Suppl 1):S7–S14.
18. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires.* New York: Oxford University Press, 1996.
19. SAS Institute. *SAS/STAT Software: Changes and Enhancements* (version 6.11). Carey, NC: SAS Institute, 1996.
20. SAS Institute. *SAS/STAT Software: SAS/STAT Users' Guide Vol 2* (4th ed). Carey, NC: SAS Institute, 1990.
21. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State — a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12:189–198.
22. Cornoni-Huntley J, Brock DB, Ostfeld AM, Taylor JO, Wallace RB. *Established Populations for Epidemiologic Studies of the Elderly: Resource Data Book.* Bethesda, MD: National Institute on Aging, 1986. NIH publication no. 86-2443.
23. Fried LP, Bandeen-Roche K, Williamson JD, et al. Functional decline in older adults: expanding methods of ascertainment. *J Gerontol Med Sci.* 1996;51A:M206–M214.
24. Guralnik JM, Simonsick EM, Ferucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *J Gerontol Med Sci.* 1994;49:M85–M94.
25. Guralnik JM, Branch LG, Cummings SR, Curb JD. Physical performance measures in aging research. *J Gerontol Med Sci.* 1989;44: M141–M146.
26. Guralnik JM, Seeman TE, Tinetti ME, Nevitt MC, Berkman LF. Validation and use of performance measures of functioning in a non-disabled older population: MacArthur Studies of Successful Aging. *Aging Clin Exp Res.* 1994;6:410–419.
27. Fuller WA. *Measurement Error Models.* New York: John Wiley & Sons, 1987.
28. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol.* 1991;134:1233–1244.

Received March 20, 1997
Accepted December 9, 1997