American Journal of Epidemiology
Vol. 146, No. 5
Printed in U.S.A.


# Proxy Reporting in Five Areas of Functional Status

## Comparison with Self-Reports and Observations of Performance

Jay Magaziner, Sheryl Itkin Zimmerman, Ann L. Gruber-Baldini, J. Richard Hebel, and Kathleen M. Fox

Proxy ratings of functional status were compared with subject self-reports in five domains relevant to the study of older persons and with observations of subject performance in two areas (physical and instrumental functioning). Data were derived from 233 proxy-subject pairs evaluated in a prospective study of hip fracture patients aged 65 years or more in Baltimore, Maryland (1990–1991). Agreement between proxy and subject reports was highest for a summary measure of instrumental functioning and lowest for a measure of depression. Proxies tended to report more disability than did subjects, although bias varied by function. Patterns of agreement for proxy reports versus observations of performance compared with patterns for proxy reports versus subject reports were lower for measures of instrumental functioning, and bias was generally more extreme for instrumental and physical functioning measures. The authors conclude that agreement and bias differ by functional domain, by the way summary measures are created and scored, and by the criterion against which proxy reports are compared. *Am J Epidemiol* 1997;146:418–28.

activities of daily living; aged; cognition; data collection; epidemiologic methods; frail elderly; questionnaires; research design

Epidemiologic research relies heavily on information obtained directly from study subjects. Under some circumstances, however, it is necessary to obtain data about subjects from a surrogate or proxy, such as when studying older persons, where physical, sensory, and cognitive deficits may be impediments to obtaining data directly. It is not uncommon for more than 20 percent of the community-dwelling aged, 40 percent of the hospitalized aged, and 50 percent of aged nursing home residents to be unable to provide information because of their limitations (1–6).

Many studies have examined proxy responses; most have focused on areas such as diet, smoking, occupational exposure, and other chronic disease risk factors (7–21). Several studies (6, 22–27) have evaluated proxy responses to questions about physical and instrumental functioning in older persons; with few exceptions (23, 27–30), this research has not considered the accuracy of proxy reports for measures of social, affective, or cognitive functioning, all of which are important dimensions of the overall health and quality of life of older persons.

Although it is often overlooked, the criterion against which proxy ratings are compared is important. Most prior studies of the aged have compared proxy ratings with responses provided by the subjects themselves (6, 23, 25–28, 31); some (22, 24, 30, 32, 33) have compared proxy responses with medical records and judgments made by health care providers. A few studies (31, 34–36) have compared proxy responses with observations of subject performance of functional tasks following explicit protocols. While performance measures are not a gold standard, they are an important criterion against which to compare proxy reports, as they have the potential to provide a more discrete assessment of functioning than a self-report does.

The purpose of the present study was to evaluate proxy ratings in five functional domains frequently used in studies of older persons: physical, instrumental, cognitive, affective, and social functioning. Criterion measures for these comparisons included subject self-reports and observed subject performance of selected physical and instrumental tasks. The extent to which proxy ratings corresponded to measures relying on other sources (i.e., agreement) was assessed, and the degree to which proxies over- or underreported in

Received for publication August 12, 1996, and in final form April 7, 1997.

Abbreviations: CES-D, Center for Epidemiologic Studies Depression scale; IADLs, instrumental activities of daily living; ICC, intraclass correlation coefficient; MMSE, Mini-Mental State Examination; PADLs, physical activities of daily living.

From the Department of Epidemiology and Preventive Medicine, School of Medicine, University of Maryland, Baltimore, MD.

Reprint requests to Dr. Jay Magaziner, University of Maryland School of Medicine, 660 West Redwood Street, Suite 200, Baltimore, MD 21201.

comparison with these sources (i.e., bias) was evaluated.

## MATERIALS AND METHODS

### Subjects and data collection

Subjects included 233 individuals participating in the 12-month follow-up evaluation of a prospective study of hip fracture recovery, and a self-designated proxy for each. Hip fracture patients entering one of eight Baltimore, Maryland, hospitals from January 1, 1990, through June 15, 1991, who were 65 years of age or older and living in the community at the time of their fracture were identified for inclusion in the study. At the time of fracture, consent to participate in the prospective study was obtained from 674 patients; 328 (49 percent) were eligible for inclusion in the proxy component of the project, which began 5 months after subjects started to become due for their 12-month follow-up evaluation. Subjects were asked to identify the person most knowledgeable about their health and general abilities. These individuals were then asked to participate in this study as the subject's proxy respondent. Two hundred and thirty-three (71 percent) of the eligible subjects completed an interview, provided the name of a proxy who consented to participate, and had proxies who completed an interview within 1 month of the patient's 12-month evaluation. Subjects were evaluated at their place of residence (home or institution); proxies were interviewed by telephone. All evaluations were conducted by research staff who had been trained in interviewing and measuring performance in older subjects.

### Measures

Subjects were asked questions about their functional status in five areas; proxies were asked a similar series of questions, with questions rephrased to refer to the subject. Subjects also were observed performing selected tasks in two areas of functioning (physical and instrumental).

*Physical functioning.* Information was obtained about assistance used in the past week to perform 15 physical activities of daily living (PADLs), using the structure of the Functional Status Index (37). Patients and proxies were asked whether the patient had received assistance in carrying out each PADL within the past week, and responses were coded as follows: 1) performed the task independently, 2) performed the task with assistance (human help and/or equipment), 3) did not perform the task for health reasons, or 4) did not perform the task for non-health reasons (e.g., the absence of stairs precluded stair-climbing). For comparisons of proxy reports with subject self-reports, responses were dichotomized, classifying the subject as independent versus dependent (defined as needing assistance or not performing the task for health reasons). Subjects reporting that they did not perform activities for non-health reasons were eliminated from the analyses, since this type of nonperformance did not signify a lack of independence. Three PADL summary scales were created: 1) upper extremity PADLs, 2) lower extremity PADLs, and 3) a PADL summary scale incorporating functions used by Katz et al. (38). For each scale, the score was the number of items for which the subject was rated as not independent. (See the Appendix table for a listing of items and scales.)

*Instrumental functioning.* Information on instrumental activities of daily living (IADLs) was obtained using a modified version of the Older Americans Resources and Services instrument (39), which was adapted to ask subjects how they had performed seven activities during the previous 2 weeks, rather than asking them to rate their potential ability to perform activities. Responses to each item were dichotomized as independent versus dependent (used assistance or was unable to perform the task for health reasons). A summary IADL scale was created which counted the number of the seven activities in which subjects were dependent. (See the Appendix table for a listing of items.)

*Affective functioning.* The Center for Epidemiologic Studies Depression scale (CES-D) (40) was used to measure affective status. This 20-item instrument consists of questions that describe behaviors and feelings. Subjects were asked to indicate how often within the past week they had behaved or felt a certain way (rarely, sometimes, occasionally, or most of the time); proxies were asked to rate how often they had thought the subject felt this way. Scores on the CES-D range from 0 to 60, with higher scores indicating more depressive symptomatology. Scores of 16 or greater are considered indicative of significant depressive symptomatology.

*Cognitive functioning.* The Mini-Mental State Examination (MMSE) (41) was used to assess the cognitive status of subjects. It is scored on a scale ranging from 0 to 30, where scores of 23 or less are indicative of cognitive impairment. Proxies were asked to estimate how the subject would do on each of the items in the MMSE. When the MMSE was administered to proxies, multiple-task items were combined and questions were asked as one unit. For each of these items (i.e., three-stage command; repeat three items; recall three items; and serial subtraction), the proxy score was either 0 or the maximum point value. For example, the three-stage command asks subjects to take a piece of paper in their right hand, fold it in half, and

then place it on the floor; the subject can receive 0–3 points. Proxies were asked whether the subject would be able to complete the entire task; the proxy score was 0 or 3, with no intermediate values.

*Social functioning.* The 12 social functioning items included passive, active, social, and solitary activities (42). Ten of the 12 items were coded as the frequency of engaging in the activity during the past 2 weeks. Two items—watching television and reading—were coded on a 0- to 5-point scale indicating the average amount of time spent per day in each activity during the past 2 weeks (0 = 0 minutes, 1 = <15 minutes/day, 2 = 15 minutes–1 hour/day, 3 = 1–2 hours/day, 4 = 3–4 hours/day, and 5 = ≥4 hours/day). Three summary scales were formed from these 12 items: 1) social total—a scale which summed the number of the 12 activities in which the subject participated; 2) television/reading frequency—a scale for television and reading assessing the amount of time spent per day in the past 2 weeks; and 3) social frequency—a scale for the other 10 items assessing frequency of participation in the activities. (See the Appendix table for a listing of items and scales.)

*Performance-based measures of physical and instrumental functioning.* Performance for 10 of the PADLs and three of the IADLs was observed by examiners at the subject's place of residence. (See the Appendix table for a listing of items and task specifications.) For each of the items, all parts of a task had to be performed correctly and independently to obtain a rating of independent. Proxy reports and subject performance were dichotomized as independent versus dependent. Subjects who did not complete tasks correctly or who reported that they had not carried out tasks for health reasons were classified as dependent for that task.

### Data analyses

To assess comparability of responses for the categorical measures, we calculated Cohen's kappa statistic; for continuous measures, we used the intraclass correlation coefficient (ICC). These agreement statistics indicate the proportion of variance which can be attributed to between-respondent variation, as opposed to within-respondent variation (i.e., disagreement between subject and proxy). Kappa can also be interpreted as the proportion of agreement beyond the amount which is expected by chance alone. Both kappa and ICC range from less than 0 to 1, with a value of 1 indicating perfect agreement. Guidelines for deciding when agreement is less than satisfactory suggest that kappa and ICC values greater than 0.8 indicate almost perfect agreement, values between 0.6 and 0.8 indicate substantial agreement, values between 0.4 and 0.6 indicate moderate agreement, and values less than 0.4 indicate slight to fair agreement (43). These guidelines, although arbitrary, are useful for interpreting kappa and ICC values. The standard error used for kappa is that given by Fleiss et al. (44), and the standard error used for ICC is that provided by Donner and Wells (45). Agreement between proxy reports and both subject self-reports and observed performance was calculated.

In addition to agreement, analyses examined the percentage of bias in proxy ratings as compared with subject self-reports and observed performance (26). For categorical measures, percent bias was calculated as the ratio of the difference between the proportion of positive responses given by proxies and the subjects' self-reports or performance, expressed as a percentage of the proportion of subjects responding positively. For continuous measures, percent bias was determined as the difference in mean values between proxy responses and subjects' self-reports or performance, expressed as a percentage of the subjects' mean. A positive percent bias indicates that proxies reported the presence of an item more often than subjects did. Bias was tested for statistically significant departures from 0 using McNemar's chi-square test (46) for categorical measures and the paired $t$ test for continuous measures.

Comparisons of proxy reports of functioning with the functional measures provided by subject self-reports were restricted to those subjects who were not severely cognitively impaired (MMSE score >16) ($n$ = 205). Comparisons of ratings on cognitive performance were not restricted in this manner in order to maximize variability in MMSE scores.

## RESULTS

Of the 233 subjects, 78 percent were female, 59 percent were widowed, and 7 percent were nonwhite. Their average age was 80.9 years, and their mean educational level was 11.5 years (54 percent had completed high school). At the time of the 12-month assessment, 33 percent of the subjects lived alone, 50 percent lived with others in the community, and 15 percent were institutionalized. Subjects had a mean MMSE score of 24.7, with 27 percent scoring in the impaired range (≤23). The mean CES-D score was 12.3, with 28 percent scoring in the depressed range (≥16).

The average age of proxies was 60.6 years; 77 percent were female, and 65 percent were married. The mean educational level of proxies was 13.3 years (82 percent had completed high school). Proxies were most frequently children (38 percent), spouses (17 percent), or other relatives (28 percent); 38 percent of

proxies lived with the subject, and 20 percent had not had weekly contact with the subject in the month prior to the proxy interview. Proxies who did not live with subjects had had an average of 10.4 visits with the subject in the previous month; all of the proxies who did not see the subject weekly had spoken with the subject over the telephone at least once per week during the preceding month. Most proxies said they knew the subject's health status "very well" (72 percent) or "pretty well" (19 percent); only three proxies said they did not know the subject's health status.

### Proxy reports versus subject self-reports

*Agreement.* Information on the agreement and bias between proxy reports and subject self-reports is provided in tables 1 and 2. Agreement levels for the three summary measures of physical functioning (table 1) were substantial to moderate: The ICC was 0.68 for the lower extremity PADL measure, 0.56 for the upper extremity PADL measure, and 0.65 for the PADL summary scale. Levels of agreement for individual PADL items (table 2) were generally lower than those for summary measures. Of the 15 PADL items for which proxy reports were compared with subject self-reports, agreement was substantial ($\kappa \geq 0.6$) for one, moderate ($0.6 > \kappa \geq 0.4$) for six, and poor ($\kappa < 0.4$) for the remaining seven. Neither agreement nor bias could be evaluated for eating because of the low frequency of dependence on this item.

The agreement level for the IADL summary measure (table 1) was in the almost-perfect range (ICC = 0.85). Of the seven items for which proxy and subject reports were compared (table 2), almost-perfect agreement ($\kappa \geq 0.8$) was observed for two items, substantial agreement ($0.8 > \kappa \geq 0.6$) was seen for three items, moderate agreement ($0.6 > \kappa \geq 0.4$) was seen for one item, and poor agreement ($\kappa < 0.4$) was found for one item. (Compared with individual items in scales, better

**TABLE 1. Correspondence between proxy reports and subject self-reports for summary measures in five functional status domains, Baltimore, Maryland, 1991–1992**

| | No. of pairs | Mean score | | Intraclass correlation | Standard error | Bias (%) |
|---|---|---|---|---|---|---|
| | | Proxy report | Subject self-report | | | |
| Physical functioning | | | | | | |
| Lower extremity activities (11 items)† | 161 | 5.9 | 5.8 | 0.68 | 0.06 | 2.3 |
| Upper extremity activities (4 items)‡ | 189 | 0.5 | 0.3 | 0.56 | 0.06 | 77.0*** |
| Summary of six activities§ | 178 | 2.5 | 2.4 | 0.65 | 0.06 | 4.2 |
| Instrumental functioning | | | | | | |
| Summary of seven activities¶ | 137 | 3.0 | 2.9 | 0.85 | 0.05 | 5.6 |
| Social functioning | | | | | | |
| No. of activities performed in past 2 weeks (12 items) | 197 | 4.3 | 4.4 | 0.61 | 0.06 | –2.3 |
| No. of minutes per day spent reading and watching television | 176 | 6.1 | 6.4 | 0.62 | 0.06 | –5.3** |
| Frequency of participation in activities during past 2 weeks (10 items) | 189 | 9.0 | 8.5 | 0.47 | 0.06 | 6.0 |
| Cognitive functioning | | | | | | |
| Mini-Mental State Examination score | | | | | | |
| Continuous scoring | 212 | 26.3 | 25.3 | 0.65 | 0.05 | 4.2*** |
| Dichotomous scoring (% ≤23) | 212 | 21.2 | 23.1 | 0.51# | 0.06 | –8.2 |
| Affective functioning | | | | | | |
| Center for Epidemiological Studies Depression scale | | | | | | |
| Continuous scoring | 186 | 13.6 | 11.9 | 0.45 | 0.07 | 14.4* |
| Dichotomous scoring (% ≥16) | 186 | 32.8 | 29.0 | 0.38# | 0.07 | 13.0 |

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (for a test of zero bias).
† Sum of 11 lower extremity tasks for which dependent status was reported.
‡ Sum of four upper extremity tasks for which dependent status was reported.
§ A summary of dependent status in six areas of physical functioning, incorporating items used by Katz et al. (38).
¶ Sum of seven instrumental activities for which dependent status was reported.
# Kappa value.

**TABLE 2. Correspondence of independence versus dependence between proxy reports and subject self-reports and performance for individual physical and instrumental activities of daily living, Baltimore, Maryland, 1991–1992**

| | No. | Dependent (%)† | | | Proxy vs. self-report | | | Proxy vs. performance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Proxy report | Subject self-report | Subject performance | Kappa | SE‡ | Bias (%) | Kappa | SE‡ | Bias (%) |
| Physical activities of daily living | | | | | | | | | | |
| Walking 10 feet | 192 | 57.3 | 44.8 | 40.6 | 0.58 | 0.06 | 27.9*** | 0.59 | 0.06 | 41.0*** |
| Getting into and out of bed | 152 | 33.6 | 29.6 | 34.2 | 0.39 | 0.07 | 13.3 | 0.37 | 0.08 | −1.9 |
| Putting socks and shoes on | 169 | 32.0 | 32.0 | 30.8 | 0.51 | 0.07 | 0.0 | 0.59 | 0.06 | 3.8 |
| Getting on and off the toilet | 168 | 44.0 | 68.5 | 70.8 | 0.40 | 0.07 | −35.7*** | 0.31 | 0.07 | −37.8*** |
| Eating§ | 199 | 3.0 | 1.5 | 2.0 | | | | | | |
| Rising from an armless chair | 164 | 62.2 | 52.4 | 41.5 | 0.26 | 0.08 | 18.6* | 0.38 | 0.07 | 50.0*** |
| Putting on a shirt/blouse | 184 | 17.4 | 10.9 | 9.8 | 0.56 | 0.06 | 60.0** | 0.50 | 0.06 | 77.8** |
| Buttoning a shirt/blouse | 182 | 17.0 | 7.1 | 6.6 | 0.51 | 0.06 | 138.5*** | 0.31 | 0.07 | 158.3*** |
| Grooming | 190 | 14.2 | 8.4 | 2.6 | 0.35 | 0.07 | 68.8* | 0.22 | 0.07 | 440.0*** |
| Getting into and out of a bath/shower | 133 | 63.9 | 88.7 | 88.0 | 0.32 | 0.08 | −28.0*** | 0.17 | 0.09 | −27.4*** |
| Taking a shower, bath, or sponge bath | 176 | 59.7 | 43.8 | | 0.18 | 0.07 | 36.4*** | | | |
| Walking one block | 158 | 69.6 | 67.1 | | 0.77 | 0.05 | 3.8 | | | |
| Climbing five stairs | 157 | 82.2 | 93.0 | | 0.18 | 0.08 | −11.6** | | | |
| Getting into a car | 168 | 66.7 | 46.4 | | 0.15 | 0.08 | 43.6*** | | | |
| Putting on pants | 181 | 23.8 | 15.5 | | 0.43 | 0.07 | 53.6*** | | | |
| Instrumental activities of daily living | | | | | | | | | | |
| Using the telephone | 177 | 16.9 | 17.5 | 38.4 | 0.35 | 0.07 | −3.2 | 0.38 | 0.07 | −55.9* |
| Handling money | 152 | 30.9 | 34.2 | 48.7 | 0.81 | 0.05 | −9.6 | 0.42 | 0.07 | −36.5* |
| Taking medications | 166 | 30.1 | 29.5 | 47.0 | 0.81 | 0.05 | 2.0 | 0.34 | 0.07 | −35.9* |
| Getting places that are out of walking distance | 176 | 72.7 | 71.0 | | 0.79 | 0.05 | 2.4 | | | |
| Shopping | 150 | 66.7 | 60.7 | | 0.64 | 0.06 | 9.9 | | | |
| Preparing meals | 152 | 35.5 | 31.6 | | 0.65 | 0.06 | 12.5 | | | |
| Housecleaning | 139 | 80.6 | 74.8 | | 0.55 | 0.07 | 7.7 | | | |

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (for a test of zero bias).
† The dependency category includes needing assistance, not performing a task for health reasons, or being unable to perform the task completely or correctly.
‡ SE, standard error.
§ Kappa and bias were not computed because of a low prevalence of dependency.

agreement is to be expected for summary measures, because random errors associated with individual item responses will be dampened by combining items.)

For social functioning (table 1), agreement was in the substantial range for comparisons of whether the subject had engaged in any of 12 activities over the past 2 weeks (ICC = 0.61) and for the measure of the amount of time spent watching television and reading (ICC = 0.62); a lower level of agreement was seen for the measure summarizing the frequency of involvement in 10 other social activities (ICC = 0.47).

Agreement between self-reported and proxy-reported cognitive functioning on the MMSE was in the substantial-to-moderate range, depending on whether MMSE score was treated as a continuous or dichotomous measure. The ICC for the continuous measure was 0.65; the kappa value for the dichotomous measure was 0.51. Levels of agreement for affective functioning on the CES-D scale were lower than for other measures of functioning. The ICC for the continuous measure was 0.45; the kappa for the dichotomous measure was 0.38.

*Bias.* As table 1 shows, percent bias was significant only for the upper extremity PADL measure of physical functioning; a greater level of disability was reported by proxies than by subject self-reports. For other summary measures of physical functioning, proxy reporting bias was less than 5 percent and was not statistically significant. Examination of percent bias for individual PADL items (table 2) indicated that proxies generally reported more disability than subjects reported about themselves. This pattern was observed for 10 of the 14 PADL items evaluated, and was statistically significant for eight. A statistically significant pattern of underreporting of disability by proxies in comparison with subjects was seen for three items. No bias was observed for one PADL item. Proxies tended to rate more disability in instrumental functioning than did patients, although percent bias was not statistically significant. Bias for individual IADL items was generally less pronounced than for PADL items; none of the bias estimates were statistically significant.

For the social functioning summary measures, proxies reported that subjects spent less time watching television and reading per day than subjects reported for themselves. Examination of the television and reading items (not shown) revealed that percent bias on this measure was attributable to underreporting by proxies of time spent reading. Although the result was not statistically significant, proxies reported a greater frequency of participation in other activities than did patients.

Proxies significantly overrated subject cognitive

functioning (MMSE) when the scale was scored as a continuous measure. Percent bias pointed in the same direction but was not statistically significant when the scale was treated as a dichotomous measure. Proxies were more likely to report that subjects exhibited symptoms of depression (affective functioning) than subjects were to report symptoms in themselves; this association was statistically significant only when the CES-D was scored as a continuous measure.

### Proxy reports versus observations of subject performance

Table 2 provides information on level of agreement and percent bias between proxy reports and observations of subject performance for nine PADLs and three IADLs. Agreement between proxy reports and observed subject performance was slight to fair ($\kappa < 0.4$) for six PADL items and two IADL items; agreement for the remainder of the items was in the moderate range ($0.6 > \kappa \geq 0.4$). Eating was not evaluated because of the low proportion of persons who were dependent. A statistically significant percentage of bias was observed for 10 of the 12 PADL and IADL items evaluated. Proxies rated a greater need for assistance than was observed for six of the nine PADLs, and less need for assistance than was observed for all three of the IADLs. As with the comparison of proxy reports to self-reports, proxies significantly underreported the subject's need for assistance in getting on and off the toilet and in getting into and out of a bath/shower.

### Proxy reports versus subject self-reports compared with proxy reports versus observations of subject performance

The median kappa and median percent bias summarizing correspondence between proxy reports and both subject self-reports and observations of subject performance are shown in figures 1 and 2. The ranges of kappa and percent bias are also shown. Figure 1 illustrates that for PADL items, levels of agreement between proxy reports and subject self-reports are similar to those between proxy reports and performance; for IADL items, the proxy versus self-report compar-

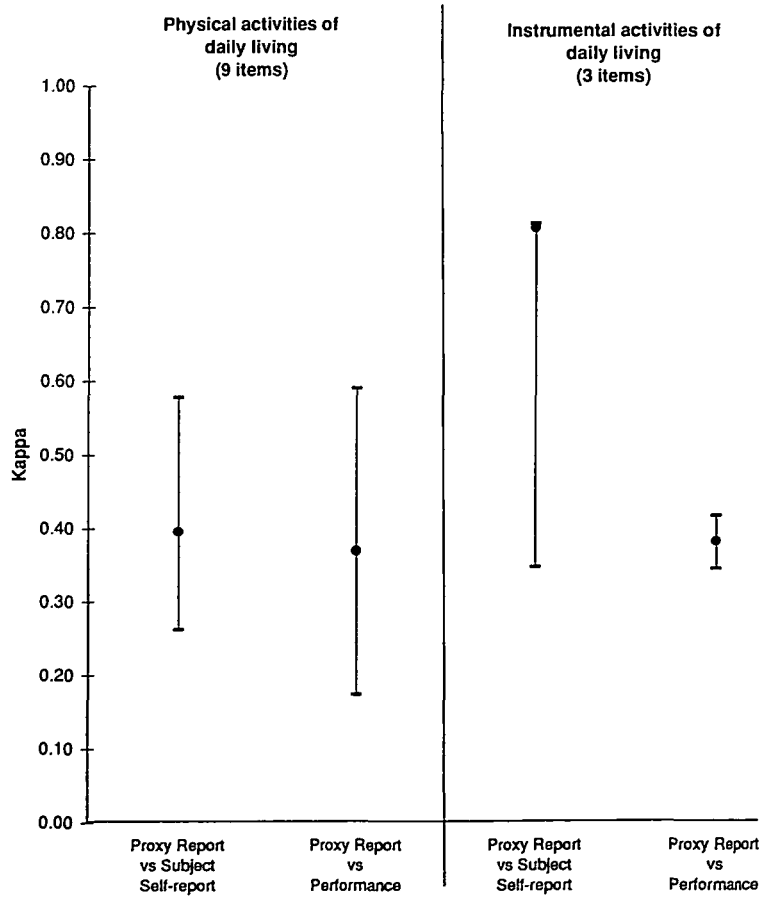


**FIGURE 1.** Median kappa values (bars, range of kappa values) for agreement between proxy reports and subject self-reports versus proxy reports and observed performance on measures of physical and instrumental activities of daily living, Baltimore, Maryland, 1990–1991.

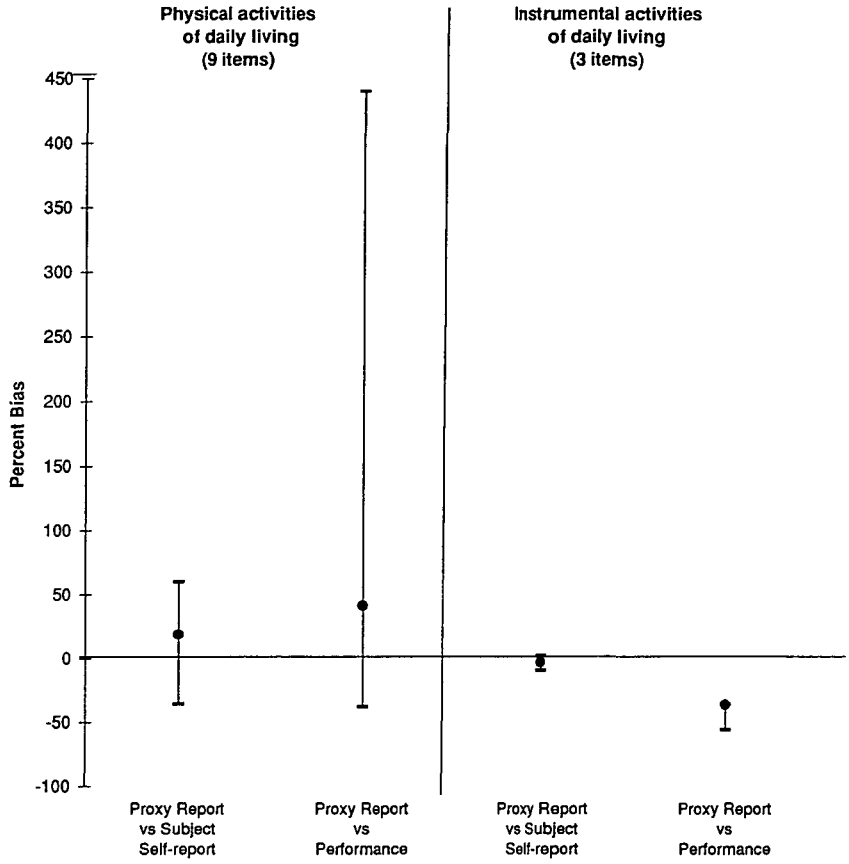


**FIGURE 2.** Median percentage of bias (bars, range of bias) between proxy reports and subject self-reports versus proxy reports and observed performance on measures of physical and instrumental activities of daily living, Baltimore, Maryland, 1990–1991.

ison has a wider range of agreement levels than does the proxy report versus performance comparison. Comparison of the percentages of bias in figure 2 indicates that the median bias for PADL items is lower for proxy reports versus self-reports than it is for proxy reports versus observed performance (18 percent vs. 41 percent). The ranges for comparisons of PADL items overlap, although two of the bias values contrasting proxy reports versus performance are greater than the upper bound of the values for proxy reports versus self-reports. Comparison of median biases and ranges of biases for the three IADL items indicates that the bias for proxy reports versus performance is negative and considerably larger than that for proxy reports versus self-reports. The median percent bias for proxy reports versus observed subject performance is −37 percent, and the median for proxy reports versus subject self-reports is −3 percent; the ranges do not overlap.

## DISCUSSION

This study examined proxy reporting for five areas of functioning that are frequently considered in health studies of older persons (47–49). Results indicate that agreement and bias differ by functional domain and specific tasks within domains, by the manner in which summary measures are created and scored, and by whether proxy reports are compared with subject self-reports or observations of performance.

### Proxy reports versus subject self-reports

In general, the more observable and less private the function being measured, the greater the agreement between proxy and subject reports. Agreement was highest for the instrumental functioning summary measure, which was based on reports of the actual performance of seven tasks, most of which are performed regularly and can be observed directly or the results of which can be readily seen by others. The lowest agreement found was for the measure of affective functioning (CES-D score), which is composed primarily of emotional states. These feelings tend to be private and would not be known by others unless articulated by the subject.

The extent and direction of reporting bias varied, with a general tendency for proxies to report more

disability than subjects reported for themselves. Exceptions were seen for cognitive functioning, where proxies attributed slightly higher levels of cognitive ability to subjects than their test performance demonstrated, and three PADL items involving possible use of handrails that may be considered to indicate dependence more by patients than by proxies: getting on and off the toilet, getting into and out of the bath/shower, and climbing five stairs.

These general patterns of concordance are consistent with previous reports of response agreement and bias (6, 22–26, 28, 32, 50, 51), with one exception. Prior studies indicate that in general, proxies report greater disability in IADLs than subjects self-report (6, 22, 24, 26, 50); in the present study, none of the IADL contrasts yielded a statistically significant estimate of bias. This difference may be due to the fact that, on the basis of earlier recommendations that questions be modified to refer to explicitly defined and observable behaviors (6, 50), the Older Americans Resources and Services IADL questions asked in this study were changed to refer to actual performance of tasks, rather than perceived ability to perform them.

The manner in which summary measures are created and scored also affects response agreement and bias. Within the physical functioning domain, agreement was lowest and bias was highest for the measure summarizing upper extremity function. Findings on social functioning suggest that questions which are relatively global and which ask simply about participation (i.e., yes/no) result in less discordance than questions asking about quantity of participation. The method used to score the MMSE and CES-D also affects concordance. For both measures, proxy-subject agreement is higher using the continuous scoring method, although the use of an impaired/unimpaired dichotomy is less likely to result in a biased estimate of impairment.

### Proxy reports versus observations of subject performance

Patterns of agreement and bias between proxy reports and observations of subject performance differed in three notable ways from comparisons between proxy reports and subject self-reports. Of the three IADL tasks for which proxy reports were compared with both subject reports and performance, levels of agreement were similar for only one (i.e., telephone use). Second, in contrast to comparisons with subject self-reports, where none of seven IADL items showed a significant bias, the pattern of proxy underreporting of disability compared with observations of performance was relatively large and was statistically significant for the three IADL items evaluated. Third, for PADL items, the degree of over- or underreporting by proxies compared with observations of performance, while generally pointing in the same direction as that for comparisons with subject self-reports, was more extreme. For most comparisons, the proxy reported more disability than the performance observations indicated. (The two exceptions, getting on/off the toilet and getting into/out of the bath/shower, corresponded to those where there also was underreporting by proxies compared with subject reports.) These results on proxy reports versus observations of performance are similar to those described by others who compared summary measures of physical functioning (35, 36) and by Elam et al. (34) in the only other study we are aware of that compared proxy reports with observations of performance on separate functions.

The fact that contrasts of proxy reports with self-reports and observations of performance resulted in different levels of agreement and bias is not surprising. Several studies comparing self-reports with observations of performance on tasks similar to those evaluated here found little agreement (35, 52–56). Previous studies have suggested that self-reports and observations of performance measure different aspects of functioning; they have differential associations with health status and contribute independently to the prediction of hospitalization, nursing home placement, and mortality (57–59). Studies contrasting family proxy reports with other data sources (e.g., health care providers, self-reports, medical records) (22, 32, 36) have also reported different levels of agreement or bias dependent on the criterion against which the proxy reports are compared.

This study had several limitations. Caution must be exercised when attempting to generalize beyond this study group. The sample consisted of a subset of patients receiving treatment for a hip fracture in one of eight hospitals in a single metropolitan area. In addition, measurements were made 1 year following a hip fracture; although most recovery in functioning had occurred by then (60, 61), proxies and subjects may have been focusing on aspects of functioning thought to be related to the fracture (i.e., lower extremity) more than would persons who had not sustained such an injury. Second, the most severely cognitively impaired subjects, often the ones for whom a proxy is most likely to be required, were excluded because they could neither follow directions for performance measures nor provide reports about themselves. Finally, although performance measures followed a standardized protocol designed to assess functioning in a manner identical to the self-report questions, the performance measures focused on discrete functional tasks; proxies may have reported on different aspects of

functions than were assessed and may have referred to functioning under different circumstances than those being observed during the examination session. Despite these potential limitations, we believe that this is the first study to have examined correspondence between proxy reports and observations of performance for this number and array of PADLs and IADLs, and to have contrasted proxy reports with subject self-reports for multiple scales and scoring methods in five functional domains.

On a practical level, these results indicate that asking questions which refer to actual performance of IADL tasks rather than asking about the perceived ability to perform them (6, 22–26) may increase agreement and reduce bias. Similar questioning strategies should be extended to other areas of functioning. Results also indicate that proxies are better at rating whether or not subjects participate in activities than they are at rating the frequency of participation, suggesting that when researchers are using proxies as a data source, preference should be given to the yes/no categorical question whenever possible. Categorical scoring may be preferable for cognition and affect, as well. The few studies that have evaluated proxy measures of affect and cognition (23, 28, 30) made contrasts across the total continuum of scale scores; in this study, it was evident that although this approach results in higher levels of response agreement, when scores provided by proxies are categorized by level of impairment, bias is not statistically significant.

The objective of studying proxy reports is primarily to determine the extent to which data from proxies can be used in place of data from other sources. To date, research in this area has demonstrated that proxy response agreement and bias are a function of the question being asked, of subject characteristics, and of proxy characteristics (50, 51). The present study demonstrates that proxy response agreement and bias also vary as a function of the criterion against which proxy reports are compared. With the increasing interest in the well-being of the oldest and frailest members of the population, many of whom cannot provide information about themselves, a greater understanding of alternative sources of information on these people's health and functioning is needed.

## ACKNOWLEDGMENTS

Support for this research was provided by grants R01AG09902, R01AG06322, and R01HD0073 from the National Institutes of Health.

The authors acknowledge the participation of the following hospitals in the Baltimore Hip Fracture Studies: Franklin Square Hospital Center; Greater Baltimore Medical Center; Northwest Medical Center; St. Agnes Hospital; St. Joseph Hospital; Sinai Hospital of Baltimore; Union Memorial Hospital; and University of Maryland Medical Systems.

The authors thank Yvonne Aro and Suzanne Miller for manuscript preparation.

## REFERENCES

1. Kay DW, Bergman K. The epidemiology of mental disorders among the aged in the community. In: Birren JE, Sloane RB, eds. Handbook of mental health and aging. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1980:34–56.
2. Regier DA, Boyd JH, Burke JD Jr, et al. One-month prevalence of mental disorders in the United States: based on five Epidemiologic Catchment Area sites. Arch Gen Psychiatry 1988;45:977–86.
3. Kelsey JL, O'Brien LA, Grisso JA, et al. Issues in carrying out epidemiologic research in the elderly. Am J Epidemiol 1989; 130:857–66.
4. Cornoni-Huntley J, Brock DB, Ostfeld AM, et al. Established Populations for Epidemiologic Studies of the Elderly: resource data book. Bethesda, MD: National Institute on Aging, 1986. (NIH publication no. 86–2443).
5. Hing E, Sekscenski E, Strahan G. The National Nursing Home Survey: 1985 summary for the United States. Hyattsville, MD: National Center for Health Statistics, 1989. (Vital and Health Statistics, series 13, no. 97) (DHHS publication no. (PHS) 89–1758).
6. Magaziner J, Simonsick E, Kashner TM, et al. Patient-proxy response comparability on measures of patient health and functional status. J Clin Epidemiol 1988;41:1065–74.
7. Herrmann N. Retrospective information from questionnaires. I. Comparability of primary respondents and their next-of-kin. Am J Epidemiol 1985;121:937–47.
8. Heyman A, Wilkinson WE, Stafford JA, et al. Alzheimer's disease: a study of epidemiological aspects. Ann Neurol 1984; 15:335–41.
9. Humble CG, Samet JM, Skipper BE. Comparison of self- and surrogate-reported dietary information. Am J Epidemiol 1984; 119:86–98.
10. Lerchen ML, Samet JM. An assessment of the validity of questionnaire responses provided by a surviving spouse. Am J Epidemiol 1986;123:481–9.
11. Marshall J, Priore R, Haughey B, et al. Spouse-subject interviews and the reliability of diet studies. Am J Epidemiol 1980;112:675–83.
12. Kolonel LN, Hirohata T, Nomura AM. Adequacy of survey data collected from substitute respondents. Am J Epidemiol 1977;106:476–84.
13. Pickle LW, Brown LM, Blot WJ. Information available from surrogate respondents in case-control interview studies. Am J Epidemiol 1983;118:99–108.
14. Enterline PE, Capt KG. A validation of information provided by household respondents in health survey. Am J Public Health 1959;49:205–12.
15. Greenberg ER, Rosner B, Hennekens C, et al. An investigation of bias in a study of nuclear shipyard workers. Am J Epidemiol 1985;121:301–8.
16. Walker AM, Velema JP, Robins JM. Analysis of case-control data derived in part from proxy respondents. Am J Epidemiol 1988;127:905–14.
17. Rocca WA, Fratiglioni L, Bracco L, et al. The use of surrogate respondents to obtain questionnaire data in case-control studies of neurologic diseases. J Chronic Dis 1986;39:907–12.
18. Davanipour Z, Alter M, Sobel E, et al. A case-control study of Creutzfeldt-Jakob disease: dietary risk factors. Am J Epidemiol 1985;122:443–51.
19. Graham P, Jackson R. Primary versus proxy respondents:

comparability of questionnaire data on alcohol consumption. Am J Epidemiol 1993;138:443–52.
20. Hatch MC, Misra D, Kabat GC, et al. Proxy respondents in reproductive research: a comparison of self- and partner-reported data. Am J Epidemiol 1991;133:826–31.
21. Chandra V, Philipose V, Bell PA, et al. Case-control study of late onset "probable Alzheimer's disease." Neurology 1987; 37:1295–300.
22. Rubenstein LZ, Schairer C, Wieland GD, et al. Systematic biases in functional status assessment of elderly adults: effects of different data sources. J Gerontol 1984;39:686–91.
23. Epstein AM, Hall JA, Tognetti J, et al. Using proxies to evaluate quality of life: Can they provide valid information about patients' health status and satisfaction with medical care? Med Care 1989;27(suppl):S91–8.
24. Weinberger M, Samsa GP, Schmader K, et al. Comparing proxy and patients' perceptions of patients' functional status: results from an outpatient geriatric clinic. J Am Geriatr Soc 1992;40:585–8.
25. Kiyak HA, Teri L, Borson S. Physical and functional health assessment in normal aging and in Alzheimer's disease: self-reports vs family reports. Gerontologist 1994;34:324–30.
26. Magaziner J, Bassett SS, Hebel JR, et al. Use of proxies to measure health and functional status in epidemiologic studies of community-dwelling women aged 65 years and older. Am J Epidemiol 1996;143:283–92.
27. Rothman ML, Hedrick SC, Bulcroft KA, et al. The validity of proxy-generated scores as measures of patient health status. Med Care 1991;29:115–24.
28. Bassett SS, Magaziner J, Hebel JR. Reliability of proxy response on mental health indices for aged, community-dwelling women. Psychol Aging 1990;5:127–32.
29. McCusker J, Stoddard AM. Use of a surrogate for the Sickness Impact Profile. Med Care 1984;22:789–95.
30. Teresi JA, Golden RR, Gurland BJ, et al. Construct validity of indicator-scales developed from the Comprehensive Assessment and Referral Evaluation interview schedule. J Gerontol 1984;39:147–57.
31. Little AG, Hemsley DR, Volans PJ, et al. The relationship between alternative assessments of self-care ability in the elderly. Br J Clin Psychol 1986;25:51–9.
32. Magaziner J, Hebel JR, Warren JW. The use of proxy responses for aged patients in long-term care settings. Compr Gerontol (B) 1987;1:118–21.
33. Rozenbilds U, Goldney RD, Gilchrist PN, et al. Assessment by relatives of elderly patients with psychiatric illness. Psychol Rep 1986;58:795–801.
34. Elam JT, Graney MJ, Beaver T, et al. Comparison of subjective ratings of function with observed functional ability of frail older persons. Am J Public Health 1991;81:1127–30.
35. Kuriansky JB, Gurland BJ, Fleiss JL. The assessment of self-care capacity in geriatric psychiatric patients by objective and subjective methods. J Clin Psychol 1976;32:95–102.
36. Dorevitch MI, Cossar RM, Bailey FJ, et al. The accuracy of self and informant ratings of physical functional capacity in the elderly. J Clin Epidemiol 1992;45:791–8.
37. Jette AM. Functional Status Index: reliability of a chronic disease evaluation instrument. Arch Phys Med Rehabil 1980; 61:395–401.
38. Katz D, Ford AB, Moskowitz RW, et al. Studies of illness in the aged. The ADL index: a standardized measure of biological and psychological function. JAMA 1963;185:914–19.
39. Fillenbaum GG. Multidimensional functional assessment of older adults: the Duke Older Americans Resources and Services procedures. Hillside, NJ: Lawrence Erlbaum Associates, 1988.
40. Radloff LS. The CES-D Scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1:385–401.
41. 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;12:189–98.
42. House JS, Robbins C, Metzner HL. The association of social relationships and activities with mortality: prospective evidence from the Tecumseh Community Health Study. Am J Epidemiol 1982;116:123–40.
43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
44. Fleiss JL, Cohen J, Everitt BS. Large sample errors of kappa and weighted kappa. Psychol Bull 1969;72:323–7.
45. Donner A, Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. Biometrics 1986;42:401–12.
46. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7.
47. Fitti JE, Kovar MG. The supplement on aging to the 1984 National Health Interview Study. Washington, DC: National Center for Health Statistics, 1987. (Vital and Health Statistics, series 1, no. 21) (DHHS publication no. (PHS) 87–1323).
48. Cornoni-Huntley J, Ostfeld AM, Taylor JO, et al. Established Populations for Epidemiologic Studies of the Elderly: study design and methodology. Aging (Milano) 1993;5:27–37.
49. Eaton WW, Kessler LJ, eds. Epidemiologic field methods in psychiatry: the NIMH Epidemiologic Catchment Area Program. Orlando, FL: Academic Press, Inc, 1985.
50. Magaziner J. The use of proxy respondents in health studies of the aged. In: Wallace RB, Woolson RF, eds. The epidemiologic study of the elderly. New York, NY: Oxford University Press, 1992:120–9.
51. Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992;45:743–60.
52. Skurla E, Rogers JC, Sunderland T. Direct assessment of activities of daily living in Alzheimer's disease: a controlled study. J Am Geriatr Soc 1988;36:97–103.
53. Tinetti ME, Ginter SF. Identifying mobility dysfunctions in elderly patients: standard neuromuscular examination or direct assessment? JAMA 1988:1190–3.
54. Sager MA, Dunham NC, Schwantes A, et al. Measurement of activities of daily living in hospitalized elderly: a comparison of self-report and performance-based methods. J Am Geriatr Soc 1992;40:457–62.
55. Zimmerman SI, Magaziner J. Methodological issues in measuring the functional status of cognitively impaired nursing home residents: the use of proxies and performance-based measures. Alzheimer Dis Assoc Disord 1994;8(suppl 1): S281–90.
56. Reuben DB, Valle LA, Hays RD, et al. Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. J Am Geriatr Soc 1995;43:17–23.
57. Rozzini R, Frisoni GB, Bianchetti A, et al. Physical Performance Test and Activities of Daily Living scales in the assessment of health status in elderly people. J Am Geriatr Soc 1993;41:1109–13.
58. Guralnik JM, Simonsick EM, Ferrucci 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 1994;49: M85–94.
59. Reuben DB, Siu AL, Kimpau S. The predictive validity of self-report and performance-based measures of function and health. J Gerontol 1992;47:M106–10.
60. Magaziner J, Simonsick EM, Kashner TM, et al. Predictors of functional recovery one year following hospital discharge for hip fracture: a prospective study. J Gerontol 1990;45:M101–7.
61. Jette AM, Harris BA, Cleary PD, et al. Functional recovery after hip fracture. Arch Phys Med Rehab 1987;68:735–40.

**APPENDIX TABLE. Tasks included in various measures of physical, instrumental, and social functioning and specifications for use of various items in the performance measures**

| Functional domain and tasks | Specific functioning scale of which item is a component | Specifications for tasks in performance-based measure* |
|---|---|---|
| Physical functioning | | |
| Walking 10 feet | Lower†, Katz‡ | Walking a 3-m course at usual speed |
| Getting into/out of bed | Lower, Katz | Getting into and out of bed using subject's own bed |
| Putting socks and shoes on both feet | Lower, Katz | Putting socks and shoes on both feet using a standardized type of sock and shoe |
| Getting on/off the toilet | Lower, Katz | Getting on and off the toilet using the subject's own bathroom |
| Eating | Upper§, Katz | Feeding oneself a spoonful of cereal |
| Rising from an armless chair | Lower | Rising without using arms (use of arms was considered assistance) |
| Putting on a shirt | Upper, Katz | Putting both arms into a standardized shirt |
| Buttoning a shirt | Upper, Katz | Fastening one button |
| Grooming | Upper, Katz | Brushing hair |
| Getting into/out of bath/shower | Lower | Getting into and out of a bathtub or shower using the subject's own bathroom |
| Taking a shower/bath/sponge bath | Lower, Katz | N/A¶ |
| Walking one block | Lower | N/A |
| Climbing five stairs | Lower | N/A |
| Getting into a car | Lower | N/A |
| Putting on pants | Lower, Katz | N/A |
| Instrumental functioning | | |
| Using the telephone | IADL¶,# | Looking up a number in a standardized telephone directory and dialing the number using the subject's own telephone |
| Handling money | IADL | Counting a specific amount of money, writing a check, and balancing a checkbook |
| Taking medications | IADL | Taking two candy pills out of a bottle (proper dosage) after a timer rings (proper timing) and placing them in mouth |
| Getting places that are out of walking distance | IADL | N/A |
| Shopping for groceries or clothes | IADL | N/A |
| Preparing meals | IADL | N/A |
| Housecleaning | IADL | N/A |
| Social functioning | | |
| Watching television | TV/reading** | N/A |
| Reading | TV/reading | N/A |
| Going to religious services | Social frequency†† | N/A |
| Attending meetings | Social frequency | N/A |
| Participating in sports | Social frequency | N/A |
| Going to movies | Social frequency | N/A |
| Going to museums | Social frequency | N/A |
| Working at a hobby | Social frequency | N/A |
| Playing cards | Social frequency | N/A |
| Going on pleasure drives | Social frequency | N/A |
| Going to a family member's or friend's home for a meal | Social frequency | N/A |
| Doing volunteer work | Social frequency | N/A |

* All parts of a task must have been performed correctly and independently for the subject to be considered independent with regard to a given task.
† Task on lower extremity function scale.
‡ Task on scale incorporating functions used by Katz et al. (38).
§ Task on upper extremity function scale.
¶ NA, not applicable; IADL, instrumental activities of daily living.
# Task on instrumental activities of daily living scale.
** Task on television/reading scale.
†† Task on frequency of participation in social activities scale.