Published in final edited form as:

J Aging Health. 2017 March ; 29(2): 289–309. doi:10.1177/0898264316635566.

Measuring Physical Capacity: An Assessment of a Composite Measure Using Self-Report and Performance-Based Items

Judith D. Kasper, PhD¹, Kitty S. Chan, PhD¹, and Vicki A. Freedman, PhD²

¹Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

²University of Michigan, Ann Arbor, USA

Abstract

Objective—The objective of this study was to develop and assess a composite measure of physical capacity using self-report and physical performance items.

Method—Item response theory (IRT) is used to evaluate measurement properties of self-report and performance items and to develop a composite measure for 7,609 participants in the National Health and Aging Trends Study.

Results—Self-reports distinguish differences at the lower end of physical capacity but not at mid-to-high levels. Performance-based measures discriminate across a fuller spectrum. An IRT-based composite score, drawing on both, provides increased measurement precision across the physical capacity spectrum and detects age group differences if either self-report or performance does so—suggesting it is better suited for studying age-related changes than either measure alone.

Discussion—Self-report and performance measures have different strengths on the physical capacity spectrum. IRT provides a means of combining these different measurement approaches for analyses of physical capacity across a broad range of functioning in later life.

Keywords

physical capacity; physical function; measurement; performance tests; self-reported function

Physical capacity has long been recognized as a critical link between diseases or impairments and activity or task limitations in theoretical models of disability onset and progression (Nagi, 1965, 1976; Verbrugge & Jette, 1994; World Health Organization, 2002), as well as observational studies of older adults (Guralnik & Ferrucci, 2009; Guralnik et al., 2000; Guralnik, Ferrucci, Simonsick, Salive, & Wallace, 1995; Guralnik et al., 1994). When conceptualized at the person-level (rather than impairments in particular body functions or structures), physical capacity—embodying concepts such as strength, range of motion, stamina, and balance—is a key pathway through which individuals maintain their ability to carry out activities. Reduced or impaired capacity, due to disease or injury, contributes to activity limitations, and as suggested in Freedman's (2009) disability framework, it may

Reprints and permissions: sagepub.com/journalsPermissions.nav

Corresponding Author: Judith D. Kasper, Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway Room 641, Baltimore, MD 21205, USA. jkasper@jhsph.edu.

trigger accommodations to compensate for gaps between capacity and the ability to carry out activities.

Measurement of an individual's physical capacity has commonly relied on self-report of ability to do, and difficulty in doing, discrete functions that are the building blocks of activities, such as reaching or grasping for objects, lifting, climbing stairs, or walking distances. For several decades, these measures that were developed by Nagi (1965, 1976) to study work-related limitations have been widely used in studies of older adults. More recently, standardized tests of physical performance have been incorporated into national and community-based studies to assess physical capacity (Crimmins et al., 2008; Gill, Williams, & Tinetti, 1995; Kasper, Freedman, & Niefeld, 2013; Life Study Investigators et al., 2006). The resulting measures draw upon observation-based assessments of performance of discrete actions, such as rising from a chair or walking and reflect gradations of performance such as strength or speed. Perhaps the most widely used performance-based scale, the Short Physical Performance Battery focuses on lower body function, combining information on balance, walking speed, and timed repeated chair stands (Guralnik et al., 1994).

Although used extensively in research, self-reports provided by a study participant (or by a proxy on behalf of the participant) about one's ability to carry out discrete actions and the difficulty involved have recognized short-comings. Questions typically ask about ability to do an activity (e.g., walk three blocks) without regard to whether the person currently does the activity or has done it recently. In addition, as noted by Rejeski, Ip, Marsh, and Barnard (2010), important contextual information (e.g., how long is a block?) is usually not provided. Questions also are typically framed as ability without the use of aids or special equipment (e.g., ability to walk three blocks without the use of aids or special equipment). This wording requires speculation on the part of persons who rely exclusively on such assistance to do an activity (e.g., never walk three blocks without using a cane; Freedman et al., 2011). Prior research also suggests that self-reports of ability are influenced by self-efficacy (Seeman, Unger, McAvay, & Mendes de Leon, 1999), personality (Kempen, Steverink, Ormel, & Deeg, 1996), and other characteristics, in addition to underlying physical capabilities. Glass (1998) suggested that the influence of such characteristics is greater on self-reports of activity performance than on performance-based assessments.

Performance-based assessments offer the advantage of direct observation, but ensuring standardized assessment particularly in large population-based studies requires detailed protocols for administration, extensive interviewer training on protocols, test equipment, and development of criteria for inclusion/exclusion of study participants in tests. There can be considerable missing data for performance-based assessments for both health-related (e.g., exclusion criteria, safety concerns) and other reasons (e.g., insufficient room to conduct walking test).

As noted, although many large-scale population-based studies now incorporate both Nagi-type self-report items and performance-based assessments, there are few evaluations of these alternative approaches to measuring physical capacity. Studies using small samples to compare various self-report and performance measures have found evidence for both strong

(Alexander et al., 2000; Latham et al., 2008) and weak associations (Reuben, Valle, Hays, & Siu, 1995) and potentially important differences in correlates (Bean, Olveczky, Kiely, LaRose, & Jette, 2011). Using the Women's Health and Aging Study, a large representative community sample of older women with moderate to severe disability, Simonsick et al. (2001) developed summary scales for upper and lower extremity function from self-report measures, combining activities of daily living, instrumental activities of daily living and Nagi items, and found strong associations with performance-based measures. The association was strongest between the lower extremity self-report scale and the lower extremity performance measures. Guralnik et al. (1994) using data from the Established Populations for the Study of the Elderly compared tests that form the Short Physical Performance Battery—balance, walking speed, and repeated chair stands—with self-reported help and difficulty with mobility-related activities of daily living and with Nagi items. They found a strong association between lower extremity performance measures and self-reported measures but concluded, based on differences in adverse outcomes among persons scoring at the high end of a performance-based summary scale, that performance measures had value across a broader functional spectrum. Rejeski et al. (2010) and Rejeski et al. (2013) have argued in studies focused on mobility that self-report and performance measures offer different and unique information about capacity despite shared variance. In a study of stair climbing performance over time among older people with knee pain and low leg strength, ability to climb stairs declined more among those with low self-reported confidence in ability at baseline (Rejeski, Craven, Ettinger, McFarlane, & Shumaker, 1996).

We use the National Health and Aging Trends Study (NHATS), a nationally representative longitudinal study of people 65 years and older, to first examine the measurement of physical capacity using self-report and performance-based assessments independently, including the level of physical capacity, whether high or low, where each measure provides the most precise measurement. We then investigate whether a measurement approach that combines these indicators offers advantages over using either one in isolation. We draw upon item response theory (IRT) methods to evaluate the information provided by individual items in the self-report and performance test batteries. We then evaluate measurement precision using self-report, performance, and a combined measure of physical capacity. Finally, we address the analytic implications for detecting differences in physical capacity by examining age and race/gender group comparisons using the three approaches.

Method

Study Sample

The NHATS is a nationally representative study of persons 65 years and older drawn from the Medicare enrollment files (Kasper & Freedman, 2015; Montaquila, Freedman, Edwards, & Kasper, 2012). Data were collected on 8,245 persons in 2011, the first year of the study. In this analysis, we draw upon the subset of 7,609 persons who completed the in-person interview, all of whom lived in the community or in residential care settings other than nursing homes. On average, the in-person interview took 2 hr, with about 25 min of the interview allocated to conducting performance-based measures of physical capacity. In all, 92% (n = 7,026) of interviews were conducted with self-respondents and 8% (n = 583) with

proxy respondents. When proxy interviews were conducted, self-report items were asked of the proxy on behalf of the study participant. If the participant was eligible for any of the performance tests, the proxy was asked whether the test could be administered. Among proxy interviews, 55% ($n = 321$) of sample persons attempted at least one test and 20% attempted all five. The NHATS protocol was reviewed and approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board.

Self-Reported Measures of Physical Capacity

Rather than relying on difficulty levels (e.g., a little, some, a lot), NHATS assessed ability to carry out pairs of tasks: walking three and six blocks, bending over and kneeling down, going up 10 and 20 stairs, reaching overhead and putting a heavy book on a shelf overhead, grasping small objects, and opening a sealed jar. Respondents were first asked about ability (yes/no) to do the more challenging task without help or use of devices. Persons who said they were not able to do the more challenging version (or who answered don't know or refused) were then asked about the ability to do the less challenging one (e.g., able to walk six blocks/three blocks). Inclusion of the more challenging activities was intended to provide information across a broader spectrum of physical capacity (Freedman et al., 2011). Because paired items were hierarchical, in this analysis each of the six pairs were coded as 2 if able to do the more challenging task; 1 if answered not able, refused, or don't know to the more challenging task but able to do the less challenging task; and 0 if answered not able, refused, or don't know to the more challenging task and answered not able to do the less challenging task. Persons who responded don't know or refused to both the challenging and less challenging items were considered missing.

Performance-Based Measures of Physical Capacity

Five performance-based assessments were included in NHATS: balance (tandem, semi-tandem, full tandem, and one-leg stand with eyes open and then with eyes shut; the latter one-leg stands were added to the traditional set of balance tests to expand assessment of higher functioning), usual walking speed, rapid chair stands, grip strength, and peak air flow. These assessments have been used individually as well as in composite measures of frailty (Bandein-Roche et al., 2006) and lower extremity functioning (Guralnik & Ferrucci, 2009). Detailed descriptions of the interviewer training procedures, protocols for test administration (e.g., laying out the walking course), and the booklet used to record test results are available elsewhere (NHATS, 2011).

Scores were created for each of the five tests ranging from 0 to 4; 0 being the worst score and 4 being the best. For balance, the score reflects completion of stands by difficulty level: 1 for semi-tandem stand; 2 for full tandem; 3 for one-leg stand eyes open held for up to 16 s; 4 for one-leg stand eyes open held for 16 s to 30 s. For other tests the values represent quartiles (from 1 = lowest to 4 = highest) of the weighted distribution of scores among persons who completed a test. Scores of 0 are assigned to persons who attempt but cannot complete a test and to persons who did not do a test because of health/medical exclusions or safety concerns. Persons are missing if they were eligible (not excluded) but did not perform a test because of space constraints (walking test), no appropriate chair (repeated chair

stands), or test results are missing for other reasons. (For more details on performance test scoring, see Kasper, Freedman, & Niefeld, 2013.)

Statistical Analyses

IRT assessment of performance and self-report items—To address the IRT assumption of a unidimensional construct, in this case physical capacity, we first conduct an exploratory factor analysis, using polychoric correlations appropriate for ordinal measures, of the self-reported and performance items separately and combined. Sufficient unidimensionality for IRT analysis (McHorney & Cohen, 2000) is demonstrated if the ratio of eigenvalues between the first and second factor is ≥ 4 (Reeve et al., 2007) and if factor loadings on the dominant factor are ≥ 0.40 .

We then employ IRT methods to assess measurement properties of the individual self-report and performance items, in particular how well items discriminate between levels of physical capacity and the location on the physical capacity spectrum where items provide the most information. The Graded Response Model (GRM), which is appropriate for ordinal measures, is used to estimate the discrimination (A) and location (B) parameters (Samejima, 1969). For the A parameter, higher values indicate better discrimination. The B parameters are estimated by creating binary comparison groups based on consecutive cut-points for the response categories (e.g., for performance measures, these are 0 vs. 1-4, 0-1 vs. 2-4, 0-2 vs. 3-4, and 0-3 vs. 4; for self-report items, these are 0 vs. 1-2 and 0-1 vs. 2). Each B parameter refers to the location on the capacity spectrum where the probability of being in the lower category versus the higher category is 50%. The number of location parameters is one fewer than the number of response categories (2 for self-report; 4 for performance). We also examine scoring precision (standard error for score estimates) for scales based on self-report and performance-based items separately and based on all items combined. The physical capacity scale in each case is based on 0 being equal to the group mean and each unit being equal to the sample standard deviation. We implement all IRT analyses using Multilog 7.0 (Thissen, Chen, & Bock, 2003).

Analytic implications of alternative measurement approaches—To evaluate the analytic implications of basing a physical capacity measure on performance tests or self-report items alone, or using a measure that combines all items, we rescaled all three measures to a common 0 to 100 scale. We then used these measures to examine mean differences in physical capacity by age (5 year age groups) and age group differences (between adjacent groups) for White men, Black men, White women, and Black women. For this analysis, persons missing a score on any item were dropped from the self-report ($n = 134$) and performance measures ($n = 1,500$), leaving sample sizes of 7,475 with a self-reported score from 0 to 12 and 6,109 persons with a performance-based score from 0 to 20. The IRT score is not summative but uses available information from all self-report and performance items; thus the full sample ($N = 7,609$) receives a score.

Results

Responses to Self-Report Items and Performance Tests by Age and Gender

As Table 1 shows almost all persons could be scored on the self-report items overall, and within age and gender groups. For performance tests, 90% or more of the total sample could be scored on most items. In the three oldest age groups, percentages decline for repeated chair stand, grip strength, and peak air flow but remain above 85% (with the exception of peak air flow test for those ages 90 and older). For both men and women, the percentage with scores is 90% or above except for peak air flow for women.

Unidimensionality of Physical Capacity

Exploratory factor analyses of the self-report and performance measures separately, and of all measures combined, demonstrate sufficient unidimensionality for IRT analysis. In each case, using the eigenvalue > 1.0 criterion, a single factor emerged. Furthermore, the items load on one factor representing physical capacity, with all factor loadings well above 0.40, a commonly used cut-point (Reeve et al., 2007; Table 2). The self-report item on grasping small objects/opening a sealed jar and the grip strength performance measure have the lowest factor loadings but are above 0.60.

Item-Level Discrimination and Location for Self-Report and Performance-Based Items

Drawing from a model calibration that includes all self-report and performance items, the discrimination parameters indicate that self-report items on walking, climbing stairs, and lifting discriminate best among all items (Table 3). Five of the six self-reported items have the highest discrimination parameters overall. The upper extremity measures—the grip strength performance item and the self-report of grasping or gripping—were less discriminating than lower extremity measures and have the lowest ability to discriminate among all items. The location parameters, however, suggest that the performance-based items cover a broader range of the spectrum than the self-report items. Four of the five performance-based measures detect differences in physical capacity in persons across the spectrum—well up to 1 *SD* both above and below the mean. For balance, for example, the range is from -1.39 to 1.14 (contrasting those with scores of 0 to 3 versus those with a score of 4). In contrast, the location parameters for the self-report items are, with one exception (bending or kneeling), all in the negative range, indicating measurement using these items is best at the low end of the range of physical capacity.

Figure 1 uses walking to illustrate the differences in discrimination and location between the self-report and performance-based measures. As the location parameters indicate, the self-report item primarily discriminates at low levels of physical capacity (i.e., between -1.5 and -1.0 *SDs* below the mean on the scale) but loses the ability to discriminate beyond 1.0 *SD* above the mean. Within the narrow range of physical capacity at the lower end of the spectrum that is covered by the self-report measure, the difference in probability of being in a particular category is greater using that measure than the performance measure. For example, the probability of being in the lowest category of the self-report measure is 93% at 1.0 *SD* below the mean and 63% at 0.5 *SD* below the mean, while the difference in probability of being in the lowest category on the performance measure at the same locations

is 25% and 10% respectively. The performance score, however, provides information across a much broader spectrum (i.e., between -3.0 and 3.0 SD s on the scale) than does self-report.

Measurement Precision of Estimates From Self-Report, Performance-Based, and a Combined Measure

Figure 2 shows the standard errors for score estimates of physical capacity using all items and for the performance and self-report measures separately. For estimates of physical capacity among persons within 1 SD below the mean (i.e., -1.0 to 0.0 on the x -axis), measurement precision is better with smaller standard errors for self-report items than for performance-based items. Among those at 0.5 SD or higher above the mean (e.g., >0.5 on the x -axis), however, the performance-based items produced scores with substantially smaller standard errors than the self-report items. For the combined score, measurement precision is best at the lower end of the physical capacity scale, but standard errors for estimates for persons who have higher physical capacity remain relatively low up to 1 SD above the mean. The standard error for the combined score performed as well or better than either self-report or performance measure alone across most of the physical capacity spectrum.

Age and Gender/Race Comparisons of Self-Report, Performance-Based, and Combined Measures

Figure 3 shows mean physical capacity scores (range from 0 to 100) by age group using all three measurement approaches. Mean scores decline with age as expected. The self-report mean score is higher in each age group, however, reflecting that regardless of age a high proportion of individuals reported being able to do these activities. Performance mean scores are consistently lower than self-report means. IRT mean scores (that combine performance and self-report scores into a single score) are closer to performance means in all age groups except for those 90 years or older.

Table 4 illustrates how measures differ in their ability to detect age group differences in physical capacity by race and gender. For White men, self-report does not detect differences between age groups for those ages 65 to 80 years, whereas performance detects age group differences across the spectrum, as does the combined IRT score. For White women, all measurement approaches detect differences between age groups. For Black men, the patterns are not consistent: performance measures detect differences for two age group comparisons (65-69 years vs. 70-74 years; 75-79 years vs. 80-84 years) that are not detected by self-report or IRT; all measures detect differences for 70 to 74 years versus 75 to 79 years; no measures detect differences for 80 to 84 years versus 85 to 89 years; and self-report and IRT detect differences for 85 to 89 years versus 90 years or older. For Black women, no measures detect differences for 65 to 69 years versus 70 to 74 years, and all measures do so for all other age group comparisons. In general, with the exception of two comparisons for Black men (65-69 years vs. 70-74 years; 75-79 years vs. 80-84 years), the IRT combined score is able to detect age group differences if either the performance (as for White men ages 65-80 years) or self-report (as for Black men 85-89 years vs. 90+ years) measure detects a difference.

The combined IRT score provides better measurement precision (smaller confidence intervals) as well. For example, for Black women 75 to 79 years versus 80 to 84 years, the performance measure yields an estimate of a mean difference between these age groups of 4.93 points on a scale from 0 to 100 with 95% confidence interval (CI) = [0.86, 9.01]; the self-report measure yields a difference of 10.16 with 95% CI = [3.85, 16.47]; and the IRT combined score shows a difference of 5.86 with 95% CI = [2.63, 9.11].

Discussion

Self-report and performance-based measures of physical capacity were found to tap into a common underlying construct; however, the two approaches contribute information in different ways. Self-report measures are mainly effective in making distinctions among persons at the lower end of the spectrum, those who are unable versus others. By contrast, consistent with Guralnik et al. (1994), we find that a performance-based measure consisting of both lower and upper body tests provides information across a broader range of the physical capacity spectrum, at the higher end in particular.

Building on prior research, our analysis offers several advances in physical capacity measurement. We draw upon a national sample of older people, in which both self-report and performance items were administered. Our finding that self-report measures of physical capacity have little utility in differentiating among persons in the mid- or high range of functioning, even with more challenging items incorporated, suggests the limits of this form of assessment. The measures evaluated here had three levels—able to do the harder level, unable to do the harder level but able to do the easier level, and unable to do the easier level. However, the chances of having a score indicating ability to do the harder activity were high even for persons just above the mean (e.g., persons at 0.5 *SD* above the mean had close to a 90% probability of endorsing ability to do the harder activity). Adding additional categories (e.g., levels of difficulty) would improve the ability of self-report to differentiate more broadly across the higher functioning end of the physical capacity spectrum only if these additional categories did not also cluster at the lower end of the range (at the mean or below). A previous study found high validity and reliability between the self-report items structured to assess ability to do a more challenging and less challenging item and a measure with levels of difficulty (Freedman et al., 2011), which suggests item discrimination and location might not differ substantially between the two. Adding self-report items that ask about harder levels of performance, for example, walking longer distances or climbing more flights of stairs, is a possible approach to expanding the information from self-report items across a broader range of physical capacity but concerns noted earlier about the lack of context—what constitutes a flight of stairs—and the speculative nature of self-report items for individuals who do not do them, may be even more problematic for items that try to tap higher functioning, for example, climbing three or four flights of stairs or walking 1 mile. The measure evaluated here (six items each with a more challenging and easier version) appears effective in differentiating people who have very low physical capacity from others but used alone is less useful, for example, as a means of identifying people at higher functioning levels who might be at risk of developing low physical capacity.

Our study also demonstrates that a performance-based measure is able to discriminate among levels of physical capacity across the spectrum and is particularly useful in the mid to upper ranges of capacity. At the lower end, performance-based assessments are less effective in discriminating, which may be linked to higher levels of missing scores for these assessments. Although persons who are ineligible for (health/medical reasons) or do not attempt (for safety reasons) performance tests are assigned values (of 0) and are not missing, missing data (where values could not be assigned) are more common for performance assessments than for self-report. For example, 8.0% of persons had a missing score on walking speed, compared with 0.6% for reported ability to walk six or three blocks. Inability to score an individual on a performance test can be due to environmental factors that affect test administration (e.g., not enough room for a walking test). In addition, some of this difference is due to inclusion of proxy responses in self-report data. In NHATS, when the interview was done with proxy respondents (about 7%), sample persons were still invited to do performance tests for which they were eligible. Nonetheless, using walking as an example, 19% (108) of eligible persons with proxy respondents had missing data on the walking speed performance test, whereas only one proxy respondent failed to answer the self-report item.

In addition to using IRT to investigate the measurement properties of self-report and performance measures of capacity, we developed a combined score using IRT that draws on both sets of items. This composite score has the advantage of greater measurement precision than either the self-report or performance-based scales. The composite score is more inclusive, drawing on both self-report and performance items, so should be expected to demonstrate better measurement precision than either subset alone. However, the IRT-based score also provides a better understanding of the level of the physical capacity construct where each item is most discriminating (best at assessing lower vs. higher levels). Improved measurement precision for the composite measure stems not only from the increased information available with a larger pool of items but also from the inclusion of items that perform well at different points on the physical capacity spectrum (as shown in Figure 2). Moreover, although self-report and performance-based measures do not consistently detect age group differences for gender/race groups, with few exceptions the IRT combined score is consistently able to detect age group differences if either the performance or self-report measure detects a difference, and thus may be better suited for studying age-related changes than either set of measures alone.

Our analysis has limitations. Findings regarding the limitation of self-report measures in providing information regarding the mid and higher range of physical capacity are based on the measures available for this analysis. Although we have suggested that additional response categories for existing items, for example, levels of difficulty, that reflect harder levels, likely would not have a substantial impact on this finding, we have not tested this alternative. Whether additional self-report items could expand the ability of a measure based only on self-report to provide information across a broader spectrum of physical capacity also remains an open question. Missing data on both self-report and performance items was low. For the combined measure, which draws on all available items, the IRT score will be based on all items with a score. This has advantages in that an individual is not dropped based on missing one or two items, but does mean that some individuals will be scored

based on a subset rather than all items. The illustrative comparisons of results based on the different measurement approaches were limited to age group comparisons within race/gender. Based on our analyses of measurement properties, however, we expect the measures to behave similarly with regard to detecting physical capacity differences in other comparisons. Our analysis was also limited to an evaluation of physical capacity measures at a point in time and did not consider responsiveness to change in physical capacity over time.

Overall, the IRT measure of physical capacity by providing a mechanism to combine self-report and performance tests draws on the strengths of both and is effective in discriminating physical capacity differences across the physical capacity spectrum. Future investigation of the applications of such a measure include its advantages in predicting the impact of changes at different levels of physical capacity, for example, from the highest quartile to the third quartile of scores, for subsequent adverse events such as falls or hospitalizations and the unfolding of the disablement process. To the extent that a combined measure proves effective in identifying levels of risk tied to changes in physical capacity at earlier stages, opportunities for prevention of poor outcomes and improved quality of life at older ages are expanded.

Acknowledgments

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the U.S. Department of Health and Human Services, National Institutes of Health, National Institute on Aging, U01AG032947.

References

- Alexander NB, Guire KE, Thelen DG, Ashton-Miller JA, Schultz AB, Grunawalt JC, Giordani B. Self-reported walking ability predicts functional mobility performance in frail older adults. *Journal of the American Geriatric Society*. 2000; 48:1408–1413.
- Bandeem-Roche K, Xue QL, Ferrucci L, Walston J, Guralnik JM, Chaves P, Fried LP. Phenotype of frailty: Characterization in the women's health and aging studies. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2006; 61:262–266.
- Bean JF, Olveczky DD, Kiely DK, LaRose SI, Jette AM. Performance-based versus patient-reported physical function: What are the underlying predictors? *Physical Therapy*. 2011; 91:1804–1811. [PubMed: 22003163]
- Crimmins, E., Guyer, H., Langa, K., Ofstedal, MB., Wallace, R., Weir, D. Documentation of physical measures, anthropometrics and blood pressure in the Health and Retirement Study. 2008. (HRS Documentation Report DR-011). Retrieved from <http://hrsonline.isr.umich.edu/sitedocs/userg/dr-011.pdf>
- Freedman VA. Adopting the ICF language for studying late life disability: A field of dreams? *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2009; 64:1172–1174.
- Freedman VA, Kasper JD, Cornman JC, Agree EM, Bandeem-Roche K, Mor V, Wolf DA. Validation of new measures of disability and functioning in the National Health and Aging Trends Study. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*. 2011; 66(9):S1013–S1021.

- Gill TM, Williams CS, Tinetti ME. Assessing risk for the onset of functional dependence among older adults: The role of physical performance. *Journal of the American Geriatric Society*. 1995; 43:603–609.
- Glass TA. Conjugating the “tenses” of function: Discordance among hypothetical, experimental and enacted function in older adults. *The Gerontologist*. 1998; 38:101–112. [PubMed: 9499658]
- Guralnik JM, Ferrucci L. The challenge of understanding the disablement process in older persons. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2009; 64:M1169–M1171.
- Guralnik JM, Ferrucci L, Pieper CF, Leveille SG, Markides DS, Ostir GV, Wallace RB. Lower extremity function and subsequent disability: Consistency across studies, predictive models, and value of gait speed alone compared with the Short Physical Performance Battery. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2000; 55(4):M221–M231.
- Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. *New England Journal of Medicine*. 1995; 332:556–561. [PubMed: 7838189]
- Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, Wallace RB. A Short Physical Performance Battery assessing lower extremity function: Association with self-reported disability and predication of mortality and nursing home admission. *Journal of Gerontology: Medical Sciences*. 1994; 49(2):M85–M94.
- Kasper, JD., Freedman, VA. National Health and Aging Trends Study user guide: Rounds 1, 2, 3, & 4: Final release. Johns Hopkins University School of Public Health; 2015. Available from www.nhats.org
- Kasper, JD., Freedman, VA., Niefeld, MR. Construction of performance-based summary measures of physical capacity in the National Health and Aging Trends Study. Johns Hopkins University School of Public Health; 2013. (NHATS Technical Paper No. 4) Available from www.nhats.org
- Kemper GI, Steverink N, Ormel J, Deeg DJ. The assessment of ADL among frail elderly in an interview survey: Self-report versus performance-based tests and determinants of discrepancies. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*. 1996; 51(5):P254–P260.
- Latham NK, Mehta V, Nguyen AM, Jette AM, Olarsch S, Papanicolaou D, Chandler J. Performance-based or self-report measures of physical function: Which should be used in clinical trials of hip fracture patients? *Archives of Physical Medicine Rehabilitation*. 2008; 89:2146–2155. [PubMed: 18996244]
- Life Study Investigators. Pahor M, Blair SN, Espeland M, Fielding R, Gill TM, Guralnik JM, Studenski S. Effects of a physical activity intervention on measures of physical performance: Results of the Lifestyle Interventions and Independence for Elders Pilot (LIFE-P) study. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2006; 61:1157–1165.
- McHorney CA, Cohen AS. Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care*. 2000; 38(9 Suppl.):1143–1159.
- Montaquila, J., Freedman, VA., Edwards, B., Kasper, JD. National Health and Aging Trends Study round 1 sample design and selection. Johns Hopkins University School of Public Health; 2012. (NHATS Technical Paper No. 1) Available from www.nhats.org
- Nagi, SZ. Some conceptual issues in disability and rehabilitation. In: Sussman, MB., editor. *Sociology and rehabilitation*. American Sociological Association; Washington, DC: 1965. p. 100–113.
- Nagi SZ. An epidemiology of disability among adults in the United States. *Milbank Memorial Fund Quarterly*. 1976; 6:493–508.
- National Health and Aging Trends Study. Data collection procedures: Round 1, 2011. 2011. Available from www.nhats.org
- Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Cella D, on behalf of the PROMIS Cooperative Group. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*. 2007; 45:S22–S31. [PubMed: 17443115]

- Rejeski WJ, Craven T, Ettinger WH Jr, McFarlane M, Shumaker S. Self-efficacy and pain in disability with osteoarthritis of the knee. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*. 1996; 51:P24–P29.
- Rejeski WJ, Ip EH, Marsh AP, Barnard RT. Development and validation of a video-animated tool for assessing mobility. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2010; 65:664–671.
- Rejeski WJ, Marsh AP, Chen S-H, Church T, Gill TM, Guralnik JM, for the LIFE Research Group. The MAT-sf: Clinical relevance and validity. *The Journals of Gerontology, Series A: Biological Sciences & Medical Sciences*. 2013; 68:1567–1574.
- Reuben DB, Valle IA, Hays RD, Siu AL. Measuring physical function in community-dwelling older persons: A comparison of self-administered, interviewer-administered, and performance-based measures. *Journal of the American Geriatric Society*. 1995; 43:17–23.
- Samejima F. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 34:100–114.
- Seeman TE, Unger JB, McAvay G, Mendes de Leon CF. Self-efficacy beliefs and perceived declines in functional ability: MacArthur studies of successful aging. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*. 1999; 54:214–222.
- Simonsick EM, Kasper JD, Guralnik JM, Bandeen-Roche K, Ferrucci L, Hirsch R, Fried LP. Severity of upper and lower extremity functional limitation: Scale development and validation with self-report and performance-based measures of physical function. *The Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*. 2001; 56:S10–S19.
- Thissen, D., Chen, W-H., Bock, RD. MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory [Computer software]. Scientific Software International; Lincolnwood, IL: 2003.
- Verbrugge LM, Jette AM. The disablement process. *Social Science & Medicine*. 1994; 38:1–14. [PubMed: 8146699]
- World Health Organization. Towards a common language for functioning, disability and health ICF. The international classification of functioning, disability and health. Author; Geneva, Switzerland: 2002.

Figure 1.

Comparison of walking performance test and self-report item (IRT item characteristic curves): (a) Performance Test Walking Speed Score (0,1,2,3,4); (b) Self-report Walking Item (0,1,2).

Note. The x -axis represents levels of physical capacity where 0 is the group mean. Negative values represent levels of physical capacity below the group mean in SD s and positive values represent levels of physical capacity above the group mean in SD s. The y -axis represents the probability of endorsing a response: for performance items, scores of 0, 1, 2, 3, 4 shown as 5 lines starting from the left; for self-report items, scores of 0, 1, 2 shown as 3 lines starting from the left. Examples of interpretation: for persons in the physical capacity range at -2.0 (lower functioning at 2 SD below the mean), the probability of a score of 0 on the walking speed test is about 70% and the probability of a score of 1 is about 20%; on the walking self-report item, the probability of a score of 0 is virtually 100%. For persons in the physical capacity range at 2.0 (higher functioning at 2 SD above the mean), the probability of a score of 4 is a little over 80%, a score of 3 is about 15%, and there is a less than 5% chance of a score of 2; on the walking self-report item virtually all persons have a score of 2. IRT = item response theory.

Figure 2.

Standard error of the score estimate for physical capacity using all items, performance-based items only, and self-report items only.

Note. The x -axis represents levels of physical capacity where 0 is the group mean. Negative values represent levels of physical capacity below the group mean in SD s and positive values represent levels of physical capacity above the group mean in SD s. Standard Errors are shown on the y -axis where larger values represent larger standard errors. Standard errors are lower across the physical capacity spectrum for the measure using all items, particularly at the higher end relative to self-report only and from 1 SD below and above the mean relative to performance only.

Figure 3.

Mean physical capacity based on performance tests, self-report items and IRT score by age group.

Source. National Health and Aging Trends Study (NHATS; 2011).

Note. $N = 7,609$ persons 65 years and older. Excludes nursing home residents. IRT = item response theory.

Table 1

Percent With a Score on Performance-Based Tests and Self-Report Items by Age Group and Gender.

Tests and items	Total	Age group						Gender	
		65-69 years	70-74 years	75-79 years	80-84 years	85-89 years	90+ years	Male	Female
Total (N)	7,609	1,409	1,579	1,513	1,505	953	650	3,171	4,438
Performance tests									
Balance	94.7%	96.4%	95.7%	94.9%	93.9%	92.2%	93.2%	94.5%	94.8%
Walking speed	92.0%	94.5%	92.2%	92.3%	91.2%	90.1%	90.3%	92.0%	92.0%
Repeated chair stand	91.7%	93.2%	92.9%	91.6%	91.5%	88.4%	91.1%	91.7%	91.7%
Grip strength	92.8%	96.0%	95.1%	93.2%	92.5%	88.2%	86.9%	93.9%	92.0%
Peak air flow	90.5%	94.7%	93.3%	91.5%	89.9%	84.5%	82.3%	92.1%	89.3%
Self-report items									
Walking 6 or 3 blocks	99.4%	99.9%	99.6%	99.5%	99.3%	99.3%	98.9%	99.6%	99.4%
Climbing 20 or 10 stairs	99.0%	99.6%	99.5%	99.1%	99.1%	99.0%	99.5%	99.3%	99.3%
Lifting 20 or 10 pounds	98.1%	99.9%	99.9%	99.6%	99.6%	99.2%	99.5%	99.7%	99.6%
Bending or kneeling	97.7%	99.9%	99.9%	99.5%	99.7%	99.8%	99.5%	99.8%	99.7%
Reaching or stretching	97.2%	99.7%	99.9%	99.8%	99.6%	99.4%	99.5%	99.6%	99.7%
Grasping or gripping	96.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.7%	99.9%	99.9%

Source. National Health and Aging Trends Study (NHATS; 2011).

Note. $N = 7,609$ persons 65 years and older. Excludes nursing home residents. For performance tests, persons missing a score are those who were eligible for a test but no test information was obtained (test could not be administered, e.g., not enough space for walking test; no appropriate chair; or missing for other reasons). For self-report items, persons missing a score are those with responses of don't know or refused.

Table 2

Factor Analyses Using Polychoric Correlations of Performance Tests, Self-Report Items, and All Measures Combined.

	Performance tests	Self-report items	All measures
Performance tests and self-report items	Factor loading	Factor loading	Factor loading
Total (<i>N</i>)	6,109	7,475	6,006
Performance tests			
Balance	0.750	–	0.739
Walking speed	0.771	–	0.756
Repeated chair stand	0.729	–	0.734
Grip strength	0.575	–	0.590
Peak air flow	0.648	–	0.609
Self-report items			
Walking 6 or 3 blocks	–	0.918	0.908
Climbing 20 or 10 stairs	–	0.922	0.902
Lifting 20 or 10 pounds	–	0.929	0.923
Bending or kneeling	–	0.830	0.790
Reaching or stretching	–	0.888	0.854
Grasping or gripping	–	0.662	0.624
1st eigenvalue	2.44	4.47	6.61
2nd eigenvalue	0.22	0.13	0.48
Ratio of eigenvalue (1st and 2nd factors)	11.09	34.38	13.77

Table 3

Item Parameters for Discrimination (A)^a and Location (B)^b for Performance Tests and Self-Report Items.

Performance tests and self-report items	A	B1	B2	B3	B4
Performance tests					
Balance	2.16	-1.34	-.32	.41	1.18
Walking speed	2.25	-1.48	-.20	.52	1.24
Repeated chair stand	2.13	-.71	.09	.68	1.34
Grip strength	1.37	-1.79	-.37	.51	1.48
Peak air flow	1.44	-3.14	-.49	.49	1.45
Self-report items					
Walking 6 or 3 blocks	4.03	-.37	-.12	—	—
Climbing 20 or 10 stairs	3.76	-.66	-.32	—	—
Lifting 20 or 10 pounds	4.10	-.68	-.28	—	—
Bending or kneeling	2.35	-.73	.44	—	—
Reaching or stretching	2.82	-1.08	-.70	—	—
Grasping or gripping	1.40	-2.49	-.70	—	—

^aDiscrimination parameters indicate an item's ability to separate individuals at different levels on the physical capacity spectrum. Larger values indicate better discrimination.

^bLocation parameters describe the region of the construct where the item performs best and the cut-points are set to one fewer than the number of response categories for the item. The scale of the *x*-axis is anchored to the sample group mean (at zero) and each unit represents one *SD* of the sample (negative if below the mean; positive if above).

Table 4

Age Group Comparisons by Race and Gender Using Physical Capacity Measures Based on Performance, Self-Report, and IRT Scores.

Age group comparisons by score type	White men		Black men		White women		Black women	
	N group 1 / N group 2	Difference in group means (95% CI for mean difference)	N group 1 / N group 2	Difference in group means (95% CI for mean difference)	N Group 1 / N group 2	Difference in group means (95% CI for mean difference)	N group 1 / N group 2	Difference in group means (95% CI for mean difference)
65-69 years vs. 70-74 years								
Performance	378/399	5.05 [2.43, 7.68] *	132/139	8.11 [3.17, 13.05] *	444/483	7.30 [5.08, 9.52] *	149/177	1.89 [-2.03, 5.82]
Self-report	433/454	0.97 [-2.31, 4.25]	152/173	1.3 [-5.23, 7.85]	489/545	4.29 [0.83, 7.75] *	182/230	3.78 [-2.46, 10.03]
IRT (combined score)	436/461	3.60 [1.38, 5.83] *	152/174	3.69 [-0.24, 7.61]	491/550	4.69 [2.81, 6.56] *	184/232	2.78 [-0.46, 6.02]
70-74 years vs. 75-79 years								
Performance	399/392	5.20 [2.60, 7.80] *	139/95	5.42 [0.01, 10.83] *	483/491	8.94 [6.77, 11.10] *	177/150	9.38 [5.47, 13.30] *
Self-report	454/444	3.24 [-0.01, 6.51]	173/127	14.78 [7.90, 21.66] *	545/570	7.46 [4.13, 10.79] *	230/189	7.31 [1.13, 13.49] *
IRT (combined score)	461/447	4.43 [2.22, 6.64] *	174/129	8.81 [4.70, 12.92] *	550/581	6.51 [4.71, 8.30] *	232/196	5.13 [1.94, 8.31] *
75-79 years vs. 80-84 years								
Performance	392/365	8.85 [6.19, 11.52] *	95/90	6.83 [0.85, 12.81] *	491/487	6.05 [3.89, 8.21] *	150/150	4.93 [0.86, 9.01] *
Self-report	444/424	3.07 [-0.25, 6.38]	127/108	-1.25 [-8.95, 6.45]	570/591	6.81 [3.55, 10.07] *	189/210	10.16 [3.85, 16.47] *
IRT (combined score)	447/431	4.50 [2.25, 6.74] *	129/112	1.29 [-3.28, 5.85]	581/606	4.66 [2.91, 6.41] *	196/215	5.86 [2.63, 9.11] *
80-84 years vs. 85-89 years								
Performance	365/216	9.12 [5.98, 12.26] *	90/38	6.85 [-1.01, 14.71]	487/301	7.90 [5.43, 10.38] *	150/63	6.43 [1.13, 11.73] *
Self-report	424/273	9.79 [6.0, 13.58] *	108/58	8.40 [-1.18, 17.98]	591/396	13.26 [9.66, 16.87] *	210/101	9.75 [2.13, 17.37] *
IRT (combined score)	431/278	8.10 [5.54, 10.66] *	112/59	5.13 [-0.55, 10.82]	606/410	7.23 [5.30, 9.16] *	215/106	5.58 [1.68, 9.47] *
85-89 years vs. 90 years or older								
Performance	216/117	10.32 [6.12, 14.52] *	38/12	8.20 [-5.25, 21.65]	301/248	7.62 [4.72, 10.51] *	63/49	7.64 [0.92, 14.36] *
Self-report	273/140	10.17 [5.09, 15.24] *	58/21	19.68 [4.70, 34.67] *	396/339	13.63 [9.52, 17.74] *	101/78	14.29 [4.80, 23.77] *
IRT (combined score)	278/144	6.60 [3.19, 10.02] *	59/21	10.81 [1.83, 19.79] *	410/351	7.09 [4.90, 9.29] *	106/82	8.08 [3.25, 12.91] *

Source. National Health and Aging Trends Study (NHATS; 2011).

Note. N = 7,609 persons 65 years and older. Excludes nursing home residents. IRT = item response theory; CI = confidence interval.

* Difference in means between age groups significant at $p < .05$.