Published in final edited form as:

Res Aging. 2015 October ; 37(7): 671–694. doi:10.1177/0164027514550834.

Development of a composite measure of physical functioning for older persons

Alden L. Gross, PhD, MHS¹, Richard N. Jones, ScD², and Sharon K. Inouye, MD, MPH³

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

²Departments of Psychiatry and Human Behavior and Neurology, Warren Alpert Medical School, Brown University, Providence, RI, USA

³Aging Brain Center, Institute for Aging Research, Hebrew SeniorLife, Boston, MA, USA, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Abstract

We scaled a measure of physical functioning to a population-based normative sample by extending self-reported basic and instrumental activities of daily living with items from the MOS SF-12. We used item response theory to place items administered to a sample of older elective surgery patients on a common metric linked to the PROMIS normative sample using published data. The summary measure for physical functioning was internally consist (Cronbach's $\alpha=0.83$), reliable across a broad range of functioning, and was moderately correlated with walking speed ($r=0.52$) and energy expenditure ($r=0.40$). Demonstrating predictive criterion validity, less impaired scores were associated with lower risk of discharge to a rehabilitation facility (OR=0.38, 95% CI: 0.22,0.66) and shorter hospital stays (IRR=0.87, 95% CI: 0.79,0.97). Our approach may facilitate direct comparison of physical functioning measures across existing and future studies using a common, population-based metric, when overlapping items with the NIH PROMIS item bank are present.

Key terms

functioning; older adults; item response theory; ADL; IADL

Introduction

Physical functioning refers to the capacity to perform activities requiring physical ability (Cella et al., 2010; Rose et al., 2014). The construct encompasses activities related to independent functioning as well as more vigorous activities such as strength and endurance. Functioning with respect to self-care can be classified into basic activities of daily living

Corresponding author: Alden L. Gross, agross14@jhu.edu; Phone: 443-287-7196; 2024 E. Monument Street, Baltimore, Maryland 21205.

Financial Disclosure: No authors claim financial conflicts of interest.

(ADL) essential for self-care, such as bathing and grooming (Katz et al., 1963), and more complex instrumental activities of daily living (IADL) needed to maintain independence in the environment, such as managing money and shopping (Lawton et al., 1969; Spector et al., 1987). Functioning is commonly assessed among older adults and plays an important role in clinical decision-making and differential diagnosis, determining service needs at the individual and community level (Tonner & Harrington, 2003), and as a key outcome in observational and intervention studies in aging populations. In the Patient Reported Outcomes Measurement Information System (PROMIS) initiative, physical functioning includes the ability to care for oneself as well as lower extremity, upper extremity, and central body activities (Rose et al., 2008).

Summary measures of physical functioning based on both self-report and objective performance have been developed in community-living older adults as well as for more impaired samples (McHorney, 2002; LaPlante, 2010; Spector et al., 1998). Lawton and Brody (1969) proposed that IADL and ADL activities can be sorted along a continuum from the basic ADLs to more complex IADLs and social behaviors. In an analogous fashion, the Medical Outcomes Study (MOS) Short Form (SF) physical component score is designed to assess functional health and ability to perform tasks independently. Physical functioning can be measured by asking questions about difficulty or health limitations performing specific ADL and IADL tasks. For example, difficulty or limitations in grocery shopping might indicate a more severe disability than difficulty with feeding. Previous studies have demonstrated empirically that ADL and IADL tasks reflect points along a common continuum (Asberg et al., 1989; Bjorner et al., 1998; Fieo et al., 2011; Fisher et al., 1997; Granger et al., 1993; Haley et al., 1994; Hays et al., 2007; Heinemann et al., 1993; Jenkinson et al., 2001; Kempen et al., 1990; Linacre et al., 1994; McHorney et al., 1997; McHorney, 2002; Raczek et al., 1998; Spector et al., 1998; Stewart et al., 1992; Tonner & Harrington, 2003; Tsuji et al., 1995). Together, different scales provide a broad range of assessment of physical functioning which is desirable in heterogeneous samples (Spector et al., 1998; Stone et al., 1990).

Nearly 100 standardized questionnaires of physical functioning ability have been developed since the 1950s (e.g., Hanman et al., 1958; Mahoney et al., 1958; McHorney, 2002; Moskowitz et al., 1957; Nagi et al., 1976; Verbrugge et al., 1994). Each measure provides its own yardstick and does not relate easily to others on a common underlying continuum. Recent national efforts, culminating in PROMIS, have endeavored to use state-of-the-art psychometric methods to create summary scores with good measurement properties. PROMIS is an initiative funded by the National Institutes of Health (NIH) (<http://nihroadmap.nih.gov>). The goal of PROMIS is to provide researchers and clinicians with flexible instruments that measure constructs reliably in a manner that reduces respondent burden. The PROMIS normative sample (N=21,133) is comprised of adults 18 and over (mean age: 53, 82% white, 52% female). PROMIS instruments are appropriate for measuring patient-reported health outcomes ranging from physical health to mental health and social well-being, with physical functioning representing a primary domain (Rose et al., 2008, 2014). Findings from studies that use PROMIS measures can be more easily compared, thus facilitating the synthesis of findings across different studies.

It has been shown that combining ADL and IADL items together into a single scale provides an enhanced range of measurement of physical functioning. The present study further integrates ADL and IADL items with physical functioning questions from the MOS SF-12 and scales the resulting factor in an existing study to the PROMIS normative sample using publicly available information. The MOS SF-12, a widely used measure (McHorney et al., 1997), presents more challenging activities which extend the range of daily activities and thus expand the range of measurement. Our overall goal was to derive a summary measure of physical functioning that synthesizes information from several common physical functioning items. We developed the measure in a sample of elective surgery patients and examined its psychometric properties. We examined validity of the new metric by describing associations with performance-based measures of physical functioning and clinical outcomes. We hypothesized that the summary measure would be highly correlated with mobility and energy expenditure but not with cognitive status. Finally, we examined predictive criterion validity against clinically important outcomes.

Research Design

Participants

Our sample consisted of the first 300 consecutively enrolled elective surgery patients recruited for the Successful AGing after Elective Surgery (SAGES) study. SAGES is a prospective cohort study of cognitive and functioning outcomes of hospitalization that began in 2010 and did not use PROMIS measures. Eligibility criteria included: scheduled for a major elective surgery, aged 70 years or older, and no evidence of dementia. The Institutional Review Board at Beth Israel Deaconess Medical Center approved the study procedures.

Indicators of physical functioning

We considered seven ADL items, seven IADL items, and four items from the MOS SF-12 for inclusion as physical functioning indicators in the present study (Table 1). For basic ADLs, participants were asked if they had any difficulty in the past month with bathing, personal grooming, dressing, feeding, getting from a bed to a chair, using the toilet, and walking across a small room with no help needed (responses: no help with no difficulty/difficulty or help needed). For IADLs, participants were asked whether they had any difficulty in the past month using the telephone, getting to places outside of walking distance, shopping for groceries, preparing meals, doing housework, taking medications, and managing money (responses: help needed/no help needed). Self-reported difficulty using the telephone was non-existent (N=0) in our sample so we excluded this item, as done in previous studies (LaPlante, 2010), as well as the PROMIS item bank. We also excluded medication use and money management because of small percentages of people endorsing these items. From the MOS SF-12, we used the four items with the highest loading on the physical component summary. These items addressed physical functioning and role limitations due to physical health problems in the course of a typical day in (1) moderate activities, (2) limitations in climbing several flights of stairs, (3) limitations with work and regular activities, and (4) limitations in the kind of work and activities that can be done (Table 1).

Statistical analysis

Estimation of physical functioning—We examined the dimensionality of the functioning items using parallel analysis with scree plots (Buja et al., 1992; Horn, 1965). We then performed a factor analysis of the 17 previously described categorical indicators of physical functioning described above. We used the statistically appropriate polychoric correlations in our confirmatory factor analysis, thereby estimating a model equivalent to a logistic graded response item response theory (IRT) model (Lord, 1953; Takane et al., 1987). IRT relates responses to items to a latent trait using probabilistic models. An equation that describes the relationship between a dichotomous item and the underlying physical functioning factor θ is:

$$P(\theta)_{ij} = F[a_j(\theta_i - b_j)] \quad (\text{Eq. 1})$$

Here, the expected probability $P(\theta)$ (which is a short hand expression for $P(u_{ij}=1|q_{ij})$ for person i on item u_j) is a function of a discrimination parameter, a_j , and an item difficulty parameter b_j , a latent variable θ , which here is physical functioning, and a non-linear linking function (F), typically logistic. The model can be easily extended to ordinal response variables. The model, estimated in Mplus version 7.11, used an Expectation-Maximization algorithm for maximum likelihood estimation with robust standard errors (Muthén & Muthén, 1998–2011). Factors scores from the model, estimated by the regression-based approach in Mplus, represent the physical functioning scale.

The model separately estimates a measurement slope for each item, which provides information about how well an item differentiates between people at different levels of physical functioning. In addition, item threshold parameters identify where along the physical functioning trait an item provides the most information. There are as many thresholds for an item as category boundaries for that item. In this study, we report item slopes and threshold parameters on a standardized metric with respect to the latent variable and the indicators, so that parameters are interpretable in correlation and z-score units, respectively. Information functions were calculated based on parameters on a normal ogive scale, which is sometimes used in IRT (Lord, 1953).

Measurement slopes and threshold parameters for the NIH PROMIS normative sample are publicly available on the PROMIS website (<http://www.assessmentcenter.net>). We linked our physical functioning scale to the metric of the NIH PROMIS Physical Function item bank (version 1.0) using two items in common between SAGES and PROMIS. The two items, from the MOS SF-12, asked about limitations in moderate activities and in climbing several flights of stairs. We fixed model parameters for these items to their values in PROMIS (which are on a logistic ogive scale), which places the metric of the latent variable on the scale of the nationally representative PROMIS sample (Wave 1, N=5,239) (Liu et al., 2009). In PROMIS, responses to both items were: Not at all, Very little, Somewhat, Quite a lot, Cannot do. In SAGES, responses to these items were: No, not limited at all, Yes – limited a little, and Yes – limited a lot. We assigned the PROMIS threshold between the Not at all and Very little response options to the first threshold in SAGES, and the PROMIS threshold between Somewhat and Quite a lot to the second threshold in SAGES. In a sensitivity analysis, we estimated factors based on only the ADL/IADL items and on only the MOS

SF-12 items, and compared their distributional properties and validity to the expanded physical functioning measure.

Psychometric properties of the physical functioning measure—We examined precision (or reliability) and internal consistency. We report precision of the measure over the range of physical functioning ability using a test information curve (Hambleton et al., 1991). Internal consistency of the scale was assessed using Cronbach's alpha (Nunnally, 1967). To judge the fit of the model to data, we examined standardized differences between empirical probabilities and model-predicted probabilities. To evaluate local independence, we also compared sample polychoric correlations with model-estimated correlations using normalized residuals, computed as the standardized difference between sample and model-estimated correlations (Bollen, 1989).

Validity of the physical functioning measure—We examined convergent validity by correlating our summary physical functioning measure with performance-based time to complete a 3.5 meter walk and self-reported energy expenditure from the Minnesota Leisure Time Activities questionnaire (Pereira et al., 1997). We examined divergent validity by correlating the physical functioning measure with the 3MS, a cognitive screening test (Teng & Chui, 1987). We expected physical functioning to be moderately to highly correlated with mobility and energy expenditure and modestly correlated with cognitive status (Cohen, 1988). To determine predictive criterion validity of the physical functioning measure, we examined its ability to predict hospital length of stay using Poisson regression and risk of discharge to a rehabilitation facility using logistic regression. We expected better physical functioning to be associated with shorter hospital stays and lower risk of discharge to a rehabilitation facility.

Results

The study sample was mostly white (95%), married (62%), female (55%), and highly educated (mean 15 years)(Table 2). Of 300 participants, 48% had more than one comorbidity and 85% were scheduled for major elective orthopedic surgery. Most (61%) reported no difficulty in any ADL or IADL item, while 89% of the sample reported some limitations in the physical functioning items from the MOS SF-12. The mean level of functioning on the MOS SF-12 physical component (mean=35.6) was relatively worse than average compared to national norms (Table 2), which is consistent with the sample of predominantly orthopedic surgery patients (84%) over 70 years of age. These patients often have substantial functional limitations due to underlying orthopedic problems.

Estimation of physical functioning

Parallel analysis suggested strong evidence for a unidimensional factor underlying the 15 physical functioning indicators (Figure 1). Figure 1 plots observed eigenvalues (connected black dots) and the expected eigenvalues expected based on random reshuffling of data. The first observed eigenvalue is above that expected by randomness, while the rest fall within or below the 90% confidence region, implying unidimensionality is sufficiently met. The eigenvalue of the first factor was 8.3, while all other eigenvalues were less than 2.0, which is

below what would be expected by chance given the permutation distribution of random eigenvalues. Standardized residuals comparing empirical and model-predicted probabilities were negligible ($z < 1.7$) for every item, so local independence was sufficiently met. The proportion of variance explained for each item was above 50% for most items (Supplemental Table 1). When we computed normalized residuals, only 1 of 105 normalized residual correlations had an absolute value greater than 2.0 (results available upon request). Thus, we considered the fit of the model to the data acceptable.

Results of the IRT analysis are in Figure 2. Overall model fit was acceptable (RMSEA: 0.06; CFI: 0.96). The polychoric correlation matrix is provided in Supplemental Table 1. All items were highly correlated with the underlying physical functioning trait, as indicated by factor loadings (Figure 2). Most item location thresholds were at the more disabled end of the physical functioning continuum. ADL and IADL items had overlapping locations: while ADLs tended to provide information at the more impaired end of the functional ability continuum, several IADLs, namely meal preparation and transportation, were also informative at the impaired end.

The physical functioning measure was scored such that high values indicate less disability. Latent traits typically have a mean of 0 and variance of 1 as a condition of model identification, but PROMIS has transformed these normal scores to T scores (mean 50, standard deviation of 10) to facilitate interpretability. A score of 50 represents an average score in the NIH PROMIS normative sample, and we implemented the same transformation in our data. The mean level of the physical functioning factor in SAGES was 42.7 (SD=5.3), suggesting the SAGES sample, which is elderly and awaiting orthopedic surgery, is less physically functional on average than the normative sample by 0.7 standard deviations. There is a ceiling effect demonstrated in the less impaired range of physical functioning, constituting 11% of the sample (Figure 3, top panel).

Also in Figure 3 are distributions of physical functioning factors that used only the ADL/IADL items (second panel) and MOS SF-12 items (third panel). By using only the ADL/IADL items, 21% of the sample (N=64 participants) were at the ceiling compared to 11% for the expanded measure. Using only the MOS SF-12 items, the ceiling was unaffected but a sizable floor effect of 16% (N=48) emerged in the sample.

Psychometric properties of the physical functioning measure

Internal consistency reliability (Cronbach's alpha) of the 15 items comprising the physical functioning measure was 0.83. Analysis of the differential amount of information in the scale over the range of physical functioning, another indication of the measure's reliability, revealed reliabilities above 0.95 between scores of 30 and 50, which is between 2 SD below the PROMIS norm and the PROMIS norm (Figure 4). This range included 87.6% of the sample.

Validity of the physical functioning measure

The physical functioning measure was moderately correlated with the timed walk ($r=0.52$) and energy expenditure ($r=0.40$), providing evidence of convergent validity. The correlation was low for cognitive function measured by the 3MS ($r=0.14$), providing evidence of

divergent validity. Corresponding correlations using a physical functioning factor based only on ADL/IADL items were similar, but a factor constructed only from MOS SF-12 physical health items was less correlated than the other factors. For example, the correlation of our expanded range physical functioning scale and gait speed is 0.52, whereas restricting to the MOS SF-12 physical functioning items the correlation is 0.46. Overall differences were small but highlight the importance of including the ADL/IADL items for the criteria we have identified (Supplemental Table 2).

We examined predictive criterion validity of the physical functioning measure using rehabilitation facility placement and hospital length of stay. As shown in Table 3, the relationship of these clinical outcomes with worse functional ability increases linearly, consistent with a dose-response relationship. Among low functioning patients in the study (those with preoperative physical functional scores <35, more than 1.5 SD below the PROMIS average), 65% (95% confidence interval, CI: 40%, 84%) were placed in a rehabilitation facility, while the proportions of patients with average (scores of 35 to 50) and high (scores of 45+) levels of physical functioning were 63% (95% CI: 56%, 69%) and 36% (95% CI: 21%, 55%), respectively. Overall, the odds of discharge to a rehabilitation facility were 62% lower for each half SD higher (less impaired) physical functioning score (95% CI: 34%, 78%). Models adjusted for age, sex, race, years of education, number of comorbidities, and 3MS score to account for potential confounding of the association between the physical functioning measure and clinical outcomes.

A similar dose-response relationship was observed between the factor score and hospital length of stay. Adjusting for covariates, the mean length of stay among patients with low (scores <35), average (scores of 35 to 50), and high (scores of 50+) levels of physical functioning were 6.6 days, 5.2 days, and 5.0 days, respectively (Table 3). In multivariable analyses, a half SD higher (less impaired) physical functioning score was associated with a 13% reduction in the daily risk of remaining hospitalized (incidence rate ratio=0.87, 95% CI: 0.79, 0.97).

In a sensitivity analysis, we compared the predictive criterion validity for our physical functioning factor with scores composed of only ADL/IADL items and only MOS SF-12 items. Although the ability of these scores to predict future placement to a rehabilitation facility remained comparable to that for the full scale, ability to predict mean hospital length of stay no longer showed a dose-response relationship between the factor score and mean hospital length of stay for either of the two component scales.

Discussion

We developed a summary measure of physical functioning using items from ADL, IADL, and MOS SF-12 questionnaires. We used publicly available measures from the NIH PROMIS physical functioning item bank to calibrate our measure to the PROMIS normative sample. This feature enabled us to describe physical functioning in a selected sample on a population-based scale. Higher scores are associated with a shorter length of hospital stay and lower risk of discharge to a rehabilitation facility. The measure using these familiar items is internally consistent and provides a reliable measure of functional ability between

2.0 standard deviations below (more impaired) and the mean of PROMIS' metric. Although the reliable range of 2 SD is robust, the limitation of the reliability to the average reflects the ceiling effect observed in the sample. We observed a ceiling because most of the items in the physical functioning measure were developed for use in a more impaired population than what was recruited in SAGES. Underscoring the utility of the expanded physical functioning measure, ceiling and floor effects were more prominent when we constructed scales using only ADL/IADL or MOS SF-12 items, respectively.

This study demonstrates that the challenge of comparing findings across studies using different measures of physical functioning can be handled, to some degree, analytically. We used a surgical sample which allowed us to examine predictive criterion validity using nursing home placement and hospital length of stay. Our approach can be used by other researchers to directly compare physical functioning in existing studies with findings from new studies using the NIH PROMIS item bank. Although more overlap is better, at least one physical functioning item shared with the PROMIS item bank is enough to apply this methodology (Jones & Fonda, 2004). Because the PROMIS physical functioning item bank was constructed using existing questionnaires, our approach may be generalizable to other studies (Rose et al., 2008). The novelty in our approach is the external scaling of items from multiple measures of physical functioning to the PROMIS normative sample, which will maximize interpretability of our results and enhance comparisons across studies. The SAGES data had available ADL, IADL, and MOS SF-12 items, but other studies may use other available items as long as some questions are in common with PROMIS measures.

The major advantage of our study is that we implemented a novel approach to score a study-specific outcome measure in a manner consistent with the PROMIS metric. Using IRT to calibrate our measure may not appreciably improve the SAGES study. However, it is an important advantage to be able to compare physical functioning in our study with the normative PROMIS metric, thus enabling a more informed understanding of the generalizability and representativeness of future study samples that use PROMIS. More broadly, calibration in IRT depends on many factors, including content span of available items, precision over the range of observed ability in a sample, and differential item measurement. With respect to item content, we included all available ADL/IADL items and relevant items from the MOS SF-12, which we believe represents the construct of interest. We established our link using the PROMIS physical functioning item bank, which includes item content that overlaps both the ADL/IADL items included and the MOS/SF-12. More items, and more overlap with the PROMIS item bank, would be more optimal. Regarding measurement precision across the range of observed physical functioning, the measure provides acceptable precision where most of the sample in SAGES performs (Figures 3 and 4). Other studies may include different types of people, but the approach we took can include other items measuring physical functioning.

Our externally scaled measure of physical functioning that combined ADL, IADL, and MOS SF-12 items demonstrated less floor and ceiling effects compared to factors using ADL/IADL or MOS SF-12 alone. Whether such a score will lead to superior measures of association in all cases needs further investigation. Explicit advantages of our approach include cross-validation and pooled analyses of multiple studies to address substantive

research questions. Scale choice for a physical functioning outcome in an individual study is dependent on local factors and individual preferences. However, future advancements in research involving physical functioning could be accelerated if findings are presented on a scale common across studies, as we have demonstrated here. Regarding potential heterogeneity in the physical functioning measure, we used item parameters from PROMIS items with identical question prompts to MOS SF12 items in our study. Parameters for the other ADL/IADL items were determined empirically. Because they are determined empirically, the items only contribute to our proposed score to the extent that they share variance with the PROMIS/SF physical functioning dependency construct.

Several caveats should be mentioned. Conceptually, the dimensionality of the PROMIS physical functioning item bank has been a source of debate in the field (Hays et al., 2007; Martin et al., 2007; Raczek et al., 1998; Rose et al., 2008; Wolfe et al., 2004). Some argue that physical functioning is a heterogeneous concept that inherently reflects not only domains reflecting muscle strength and coordination, but also cognitive flexibility, and the social and environmental context of the activity being performed. Previous empirical research also suggests items assessing upper-body and lower-body functions may be different, or that mobility is distinct from self-care. In our study, we did not have items that permitted such distinctions. Rose and colleagues (2014) previously provided sufficient evidence for unidimensionality using PROMIS physical functioning data. Those findings are consistent with those from our data. Next, our sample included a large number (11%) of participants without any self-reported difficulty or help needed. This proportion suggests our measure does not discriminate well among participants at higher levels of physical functioning, resulting in an observed ceiling in the score distribution. However, this is a substantial improvement over ceiling effects present in other measures: 39% of the sample would have been at the ceiling had we only considered simple sums of IADL and ADL items. Fieo and colleagues (2011) noted that future work is needed to expand the range of the instruments that assess physical functioning for community-living older adults by including items with more sensitivity to milder disability. Lawton and Brody (1969) proposed that social activities and behaviors lie on the same continuous trait of everyday physical functioning and could provide more information in the range of physical functioning occupied by community-living older adults (Fieo et al., 2013). Another way to expand the measurable range of physical functioning might be to include more response options to ADL/IADL items, such as no difficulty, some difficulty, a lot of difficulty, and cannot do without help. However, expanded measurement of IADLs and ADLs into the more impaired range would not dilute the ceiling among community-living participants such as those in the present study. A third limitation is that our sample size was too small to evaluate differential item functioning. Individual items should measure functional ability in the same way among different subgroups of individuals, such as those defined by characteristics like age or sex (Paz et al., 2013; Thissen et al., 1988). The power to detect a minimum clinically relevant amount of bias (odds ratio of 2.0)(Cole et al., 2000) in our sample is only 72%, and the policy of the Educational Testing Service is not to examine item bias in samples of fewer than 700 participants (Zwick, 2012). Although replication of our study in a larger sample is needed, our resulting factor demonstrated acceptable measurement precision and convergent and predictive criterion validity. Fourth, we

acknowledge that the available sample size in our study is not large for factor analysis, although in this context Comrey & Lee (1992) suggested $N=300$ is “good” and Baker (1962) reported in results from a Monte Carlo study that sample sizes of $N=120$ are sufficient to estimate item parameters with variances close to those provided by asymptotic formulas. Fifth, the response distribution on the ADL and IADL items is skewed. This may have limited our ability to identify multidimensionality in the items. We note that previous studies of ADL/IADL functioning have demonstrated unidimensionality of items related to those we used (Asberg et al., 1989; Kempen et al., 1990; LaPlante, 2010). Using a more impaired sample in which items were less skewed, Spector & Fleishman (1998) concluded a single dimension explains most of the variance among ADL/IADL. Finally, calibrating physical functioning between PROMIS and the SAGES study using our approach relies on having items in common with PROMIS. There are two items in common between SAGES and PROMIS; while having more common items is more ideal, we are limited by constraints of existing data.

Advantages of the IRT-derived physical functioning measure are its higher sensitivity to differences in physical functioning in a broad range of older adults compared to scales for just ADL or IADL, allowance for different weighting for each item in the scale, and characterization of the scale’s precision over the observed range of physical functioning. Further, physical functioning derived with IRT can be treated as an interval scale, making it ideally suited for studying longitudinal change. Most importantly, advantages of the PROMIS initiative can be applied to existing resources to facilitate common measures with which to compare physical functioning across studies.

Conclusion

We derived a summary measure from widely used physical functioning measures using published data from the NIH PROMIS physical functioning item bank to calibrate the measure to the NIH PROMIS normative sample. This linking enabled us to describe physical functioning in our study on a nationally representative scale. The measure of physical functioning using familiar items was internally consistent, provided reliable measures of functional ability across low and average levels of functioning, and demonstrated no floor effect and less of a ceiling effect than other common approaches to creating physical functioning composites. The measure demonstrated predictive criterion validity: less impaired scores on the physical functioning measure were associated with lower risk of discharge to a rehabilitation facility and shorter length of hospital stay. These outcomes demonstrated a dose-response relationship. Our approach holds the potential for broad applicability to directly compare physical functioning in new and existing studies when overlapping items with the NIH PROMIS item bank are present. Importantly, these methods can facilitate interpretation and synthesis of findings across existing and future research studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funded by the National Institute on Aging (P01AG031720, SKI). Dr. Gross is supported by a post-doctoral fellowship (T32AG023480). Dr. Inouye holds the Milton and Shirley F. Levy Family Chair. The contents do not necessarily represent views of the funding entities. Funders had no deciding roles in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

References

- Asberg KH, Sonn U. The cumulative structure of personal and instrumental ADL. A study of elderly people in a health service district. *Scandinavian Journal of Rehabilitation Medicine*. 1989; 21:171–177. [PubMed: 2631192]
- Baker FB. Empirical determination of sampling distributions of item discrimination indices and a reliability coefficient. Final Report. U.S.O.E., Contract OE-2-10-071. 1962
- Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiol*. 1998; 51:1189–1202. [PubMed: 9817137]
- Bollen, KA. Structural equations with latent variables. Wiley-Interscience; 1989.
- Buja A, Eyuboglu N. Remarks on parallel analysis. *Multivariate Behavioral Research*. 1992; 27:509–540. [PubMed: 26811132]
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Choi S. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol*. 2010; 63(11):1179–1194. [PubMed: 20685078]
- Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
- Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale. experience from the New Haven EPESE study. *J Clin Epidemiol*. 2000; 53(3):285–289. [PubMed: 10760639]
- Comrey, AL.; Lee, HB. A first course in factor analysis. Hillsdale, NJ: Erlbaum; 1992.
- Fieo R, Manly JJ, Schupf N, Stern Y. Functional status in the young-old: Establishing a working prototype of an extended-instrumental activities of daily living scale. *Journal of Gerontology: Medical Sciences*. 2013; 69(6):766–772.
- Fieo RA, Austin EJ, Starr JM, Deary IJ. Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC Geriatrics*. 2011; 11:42. [PubMed: 21846335]
- Fisher WP Jr, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. *J Outcome Meas*. 1997; 1(4):329–362. [PubMed: 9661727]
- Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the functional independence measure. *Am J Phys Med Rehabil*. 1993; 72(2):84–89. [PubMed: 8476548]
- Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol*. 1994; 47:671–684. [PubMed: 7722580]
- Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
- Hanman B. The evaluation of physical ability. *N Engl J Med*. 1958; 258:986–993. [PubMed: 13541699]
- Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. *Med Care*. 2007; 45:S32–S38. [PubMed: 17443117]
- Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger C. Relationships between impairment and physical disability as measured by the functional independence measure. *Arch Phys Med Rehabil*. 1993; 74(6):566–573. [PubMed: 8503745]

- Horn JL. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
- Jenkinson C, Fitzpatrick R, Garratt A, Peto V, Stewart-Brown S. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *J Neurol Neurosurg Psychiatr*. 2001; 71:220–224. [PubMed: 11459897]
- Jones RN, Fonda SJ. Use of an IRT-based Latent Variable Model to Link Different Forms of the CES-D from the Health and Retirement Study. *Soc Psychiatry Psychiatr Epidemiol*. 2004; 39:828–835. [PubMed: 15669664]
- Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged: the index of ADL: A standardized measure of biological and psychosocial function. *JAMA*. 1963; 185:914–919. [PubMed: 14044222]
- Kempen GI, Suurmeijer TP. The development of a hierarchical polychotomous ADL-IADL scale for noninstitutionalized elders. *The Gerontologist*. 1990; 30:497–502. [PubMed: 2394384]
- LaPlante MP. The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not. *J Gerontol B Psychol Sci Soc Sci*. 2010; 65(6):720–732. [PubMed: 20100786]
- Lawton MP, Brody EM. Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*. 1969; 9:179–186. [PubMed: 5349366]
- Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. *Arch Phys Med Rehabil*. 1994; 75(2):127–132. [PubMed: 8311667]
- Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clin Epidemiol*. 2009; 63(11):1169–1178. [PubMed: 20688473]
- Lord FM. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*. 1953; 13(4):517–549.
- Mahoney F, Wood O, Barthel DW. Rehabilitation of chronically ill patients: The influence of complications on the final goal. *South Med J*. 1958; 51:605–609. [PubMed: 13556165]
- Martin M, Kosinski M, Bjorner JB, Ware JE Jr, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. *Qual Life Res*. 2007; 16:647–660. [PubMed: 17334829]
- McHorney CA. Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics Among the Oldest Old study. *Arch Phys Med Rehabil*. 2002; 83(3):383–394. [PubMed: 11887121]
- McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol*. 1997; 50:451–461. [PubMed: 9179104]
- Moskowitz E, McCann CB. Classification of disability in the chronically ill and aging. *J Chronic Dis*. 1957; 5:342–346. [PubMed: 13405990]
- Muthén, LK.; Muthén, BO. *Mplus User's Guide*. Seventh. Los Angeles, CA: Muthén & Muthén; 1998 – 2011.
- Nagi S. An epidemiology of disability among adults in the United States. *Milbank Q*. 1976; 54:439–467.
- Nunnally, JC. *Psychometric Theory*. New York, NY: McGraw-Hill; 1967.
- Paz SH, Spritzer KL, Morales LS, Hays RD. Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Qual Life Res*. 2013; 22(7):1819–1830. [PubMed: 23124505]
- Pereira MA, Fitzgerald SJ, Gregg EW, Joswiak ML, Ryan WJ, Suminski RR, Zmuda JM. A collection of Physical Activity Questionnaires for health-related research. *Med Sci Sports Exerc*. 1997; 29(6 Suppl):S1–S205. [PubMed: 9243481]
- Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries:

- Results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*. 1998; 51:1203–1214. [PubMed: 9817138]
- Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*. 1998; 51:1203–1214. [PubMed: 9817138]
- Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol*. 2008; 61(1):17–33. [PubMed: 18083459]
- Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE Jr. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol*. 2014; 67(5):516–526. [PubMed: 24698295]
- Spector WD, Fleishman JA. Combining activities of daily living with instrumental activities of daily living to measure functional disability. *Journal of Gerontology: Social Sciences*. 1998; 53:S46–S57.
- Spector WD, Katz S, Murphy JB, Fulton JP. The hierarchical relationship between activities of daily living and instrumental activities of daily living. *Journal of Chronic Diseases*. 1987; 40:481–489. [PubMed: 3597653]
- Stewart, AL.; Kamberg, CJ. Physical functioning measures. In: Stewart, AL.; Ware, JE., Jr, editors. *Measuring Functioning and Well-Being*. Durham, NC: Duke University Press; 1992. p. 86-101.
- Stone R, Murtaugh C. The elderly population with chronic functional disability: Implications for home care eligibility. *The Gerontologist*. 1990; 30(4):491–496. [PubMed: 2144254]
- Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*. 1987; 52:393–408.
- Teng EL, Chui HC. The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry*. 1987; 48(8):314–318. [PubMed: 3611032]
- Thissen, D.; Steinberg, L.; Wainer, H. Use of item response theory in the study of group differences in trace lines. In: Wainer, H.; Braun, H., editors. *Test validity*. Hillsdale (NJ): Erlbaum; 1988. p. 147-169.
- Tonner MC, Harrington C. Nursing facility and home and community based service need criteria in the United States. *Home Health Care Services Quarterly*. 2003; 24(4):65–83. [PubMed: 14998282]
- Tsuji T, Sonoda S, Domen K, Saitoh E, Liu M, Chino N. ADL structure for stroke patients in Japan based on the functional independence measure. *Am J Phys Med Rehabil*. 1995; 74(6):432–438. [PubMed: 8534387]
- Verbrugge LM, Jette AM. The disablement process. *Soc Sci Med*. 1994; 38(1):1–14. [PubMed: 8146699]
- Wang W-C. Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*. 2004; 72(3):221–261.
- Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. *Arthritis Rheum*. 2004; 50:3296–3305. [PubMed: 15476213]
- Zwick, R. [Accessed on July 25, 2014] A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. 2012. from <http://www.ets.org/Media/Research/pdf/RR-12-08.pdf>

Figure 1. Scree Plot from Parallel Analysis: Results from SAGES (N=300)

Legend. Observed eigenvalues (connected black dots) and the eigenvalues expected based on random resampling of the original data. Random data was based on 55 reshufflings of existing data.

Figure 2. Measurement Model for the Physical Functioning Measure: Results from SAGES (N=300)

Legend. Structural equation model summarizing the IRT graded response model with 17 ordinal dependent variables. Factor loadings, quantify the correlation between underlying physical functioning and each item. Item thresholds, on a $N(0,1)$ scale, depict the point along the spectrum of physical functioning in which the item has the best discrimination. To scale the physical functioning measure to the metric set for the NIH PROMIS initiative, item factor loadings and threshold parameters for two MOS SF-12 items (limitations in moderate activities and in climbing flights of stairs) were fixed to the values in the PROMIS normative sample. Numbers in parentheses denote standard errors of factor loadings and thresholds. In the figure, item slopes and threshold parameters are reported on a standardized metric, so that parameters are interpretable in correlation and z-score units, respectively.

ADL: Activities of daily living. IADL: Independent activities of daily living. MOS SF-12: Medical Outcomes Study Standard Form 12-item questionnaire.

Figure 3. Distribution of Physical Functioning in the SAGES study (N=300)

Legend. Distribution of the physical functioning score with overlaid Normal distribution for factor scores derived using all items, ADL/IADL items only, and MOS SF-12 items only. A score of 50 is the mean level of physical functioning in the NIH/PROMIS normative sample. Higher scores indicate less impairment.

Figure 4. Precision of the Physical Functioning Score Over the Range of Physical Functioning: Results from SAGES (N=300)

Legend. The information of the physical functioning score is plotted over the range of functional ability for factor scores derived using all items, ADL/IADL items only, and MOS SF-12 items only. A score of 50 is the mean level of physical functioning in the NIH/PROMIS normative sample. The shape is consistent with a score optimized for the study of between-persons differences and longitudinal change among persons with below PROMIS average physical functioning (the vertical line at a score of 50 indicates the mean of the

PROMIS normative sample). The horizontal line at a reliability of 0.95 indicates excellent reliability across most of the observed score range.

Reliability = $1 - 1 / \text{Information} = 1 - (\text{standard error of measurement})^2$

Table 1

Physical Functioning Items Used in the Harmonization: Results from SAGES (N=300)

Question	N (percent)
Basic activities of daily living, <i>number (%) reporting no help needed with no difficulty</i>	
Do you need help with bathing?	269 (89.7)
Do you need help with personal grooming?	285 (95.0)
Do you need help with dressing?	244 (81.3)
Do you need help with feeding?	292 (97.3)
Do you need help with getting from a bed to a chair?	257 (85.7)
Do you need help using the toilet?	283 (94.3)
Do you need help walking across a small room?	258 (86.0)
Independent activities of daily living, <i>number (%) reporting no help needed</i>	
* At the present time do you use the telephone?	300 (100.0)
At the present time do you get to places out of walking distance?	279 (93.0)
At the present time do you go shopping for groceries?	268 (89.3)
At the present time do you prepare your own meals?	290 (96.7)
At the present time do you do your own housework?	247 (82.3)
* At the present time do you take your own medications?	298 (99.3)
* At the present time do you manage your own money?	293 (97.7)
Medical Outcomes Study Standard Form 12-item questionnaire	
Does your health now limit you in moderate activities?	
A lot	81 (27.0)
A little	94 (31.3)
Not at all	125 (41.7)
Does your health now limit you in climbing several flights of stairs?	
A lot	108 (36.0)
A little	93 (31.0)
Not at all	99 (33.0)
Do you have any of the following problems with your work or other regular daily activities as a result of your physical health? <i>number (%) reporting no</i>	116 (38.7)
Because of your physical health, are you now limited in the kind of work or other activities you can do? <i>number (%) reporting no</i>	89 (29.7)

* These items were dropped from the final scale due to small variance.

Table 2

Participant Characteristics in the SAGES Cohort (N=300)

Characteristic	Mean or N (SD or percent)	Observed range
Age at Surgery, <i>mean (SD)</i>	76.9 (5.0)	70.1, 92.6
Sex (Female), <i>n (%)</i>	166 (55.3)	
White, <i>n (%)</i>	284 (94.7)	
Married, <i>n (%)</i>	170 (62.3)	
Education (years), <i>mean (SD)</i>	15.0 (2.9)	5.0, 20.0
Charlson comorbidity score, <i>n (%)</i>		
None	88 (29.3)	
One	69 (23.0)	
More than one	143 (47.7)	
Surgery type, <i>n (%)</i>		
Orthopedic	253 (84.3)	
Vascular	16 (5.3)	
Gastrointestinal	31 (10.3)	
Postoperative APACHEII score, <i>mean (SD)</i>	13.9 (2.8)	7.0, 24.0
3MS score, <i>mean (SD)</i>	93.2 (5.5)	71.0, 100.0
MOS SF-12 Functional ability		
Physical composite T-score, <i>mean (SD)</i>	35.6 (10.1)	13.8, 54.8
Mental composite T-score, <i>mean (SD)</i>	50.2 (8.2)	15.6, 67.2
Physical and social functioning		
Minnesota leisure time metabolic equivalents (Kcal/week), <i>mean (SD)</i>	802.1 (949.5)	0.0, 2693.9
EPESE physical activities (number of activities), <i>mean (SD)</i>	5.9 (2.0)	0.0, 9.0
Gait speed (meters/second), <i>mean (SD)</i>	0.7 (0.3)	0.2, 1.5
Outcomes of hospitalization		
Rehabilitation facility placement at discharge, <i>n (%)</i>	175 (58.3)	
Hospital length of stay (days), <i>mean (SD)</i>	5.3 (2.5)	3.0, 26.0

SD: standard deviation. ADL: Activities of daily living. IADL: Independent activities of daily living. MOS SF-12: Medical Outcomes Study Standard Form 12-item questionnaire. EPESE: Established Populations for Epidemiologic Studies of the Elderly

Table 3

Predictive Validity of the Physical Functioning Measure (N=300)

Functional ability scaled to NIH PROMIS normative sample	Sample size (N)	Proportion released to rehabilitation facility following surgery* (percent, 95% CI)	Mean hospital length of stay (days, 95% CI)
High functioning (45+)	33	0.5 (0.4, 0.6)	4.9 (4.6, 5.3)
Average functioning (35 to 45)	130	0.7 (0.6, 0.7)	5.5 (5.1, 5.9)
Low functioning (<35)	137	0.8 (0.6, 0.9)	6.1 (5.3, 7.0)

* Includes acute, subacute, and chronic care rehabilitation facilities.

Legend. Higher (more impaired) physical functioning was associated with a lower odds of discharge to a rehabilitation facility ($p<0.001$), and a lower length of stay ($p=0.02$). Estimates are adjusted for age, sex, race, years of education, number of comorbidities, and 3MS score. We selected thresholds of 35 and 45, corresponding to 1.5 SD and 0.5 SD below the PROMIS mean of 50, to divide the sample approximately into tertiles.