Measurement Article

Measuring Activity Limitations Within the National Health and Aging Trends Study (NHATS)

Elizabeth E. Marfeo, PhD, MPH, OT,^{1,*} Pengsheng Ni, MD, MPH,² Tamra Keeney, PhD, DPT,³ and Alan Jette, PhD, PT⁴

¹Department of Occupational Therapy, Tufts University Graduate School of Arts and Sciences, Medford, Massachusetts. ²Department of Health Law, Policy & Management, Boston University School of Public Health, Massachusetts. ³Center for Gerontology and Health Care Research, Brown University School of Public Health, Providence, Rhode Island. ⁴Department of Rehabilitation Sciences and Physical Therapy, MGH Institute of Health Professions, Boston, Massachusetts.

*Address correspondence to: Elizabeth E. Marfeo, PhD, MPH, OT, Department of Occupational Therapy, Tufts University Graduate School of Arts & Sciences, 574 Boston Ave. Suite 216G, Medford, MA 02155. E-mail: Elizabeth.marfeo@tufts.edu

Received: September 11, 2018; Editorial Decision Date: January 7, 2019

Decision Editor: Rachel Pruchno, PhD

Abstract

Background and Objectives: To better understand the disablement process among older adults, improved measures of activity limitations are needed. Traditional population-level measures lack the ability to distinguish precise gradations of activity limitation and are unable to detect degrees of differences over a wide range of ability levels. Therefore, we used contemporary measurement methods to improve upon current methodologies for characterizing activity limitations within the National Health and Aging Trends Study (NHATS).

Research Design and Methods: We used the NHATS Round 1 cohort to assess the feasibility of constructing an Activity Limitations scale using Rasch item response theory methods. Factor analysis was used to develop the scale from a set of existing items in the NHATS Mobility, Self-Care, and Household Activity domains. Psychometric properties of the scale were evaluated and the scale was used to examine change in activity limitations among the sample from 2011 to 2015.

Results: Results supported an 18-item scale ($N = 7,609$). Rasch infit and outfit statistics were within acceptable range for all items (Cronbach's $\alpha = 0.95$; sample score reliability = 0.83). From 2011 to 2015, 5.88% older adults demonstrated increase in function, 15% showed decrease in function, and 78% of the sample showed no change (did not exceed $\pm$ MDC90).

Discussion and Implications: Findings demonstrate that a unidimensional, interval scale of activity limitations can be constructed using traditional survey measures nested within the NHATS. Results revealed concerns regarding ceiling effects within the current self-report items of activity limitations suggesting future work is needed to expand the range of ability currently represented in the NHATS Activity Limitation items.

Keywords: Rehabilitation, Outcomes, Epidemiology, Psychometrics

Characterization of the disablement process for adults as they age requires measures of activity limitations that can distinguish meaningful gradations change over a wide range of functional ability levels. While there is a growing number of robust, contemporary measures of activity

limitations available for use within clinical contexts, there are significantly fewer options for tracking population-level health and disability. Traditionally, the study of late-life disability and function has been conducted through survey research, using methods developed over 40 years

ago (Freedman et al., 2011). These methods rely on an individual's self-report of difficulty or dependency with performing activities of daily living such as self-care and mobility and restrictions in instrumental activities of daily living (Freedman et al., 2011). Historically, responses to survey items have been summarized into basic ordinal scales that allow researchers to estimate the prevalence of late-life activity limitations and related disability. Operationally, this approach builds upon work by Nagi (1976) by asking individuals to rate their ability to perform and difficulty in performing various activities and tasks (stair climbing, walking, lifting, etc.) (Kasper, Chan, & Freedman, 2017). Items are scored on a Likert scale and then summed to provide a raw score. A primary limitation of this approach is that use of an aggregate, summated score leads to difficulty in score interpretation, as it is assumed that all items contribute equally to the construct being measured (Buz & Cortes-Rodriguez, 2016). Furthermore, the distance between scores is unknown, as a result of the ordinal nature of Likert scales (Streiner, Norman, & Cairney, 2015). The unknown interval between scores can lead to difficulties in the ability of an instrument to detect overall magnitude of change as well as change over time (Stucki, Daltroy, Katz, Johannesson, & Liang, 1996).

Specifically, these classical measurement methods have limited ability to capture granular change in activity limitations necessary for the longitudinal study of late-life disability. Increasingly, item response theory (IRT), a contemporary measurement approach, is being used to improve the measurement of health and function (das Nair, Moreton, & Lincoln, 2011; Resnick, Galik, Dorsey, Scheve, & Gutkin, 2011; Wijers et al., 2017). IRT-based approaches allow for the transformation of ordinal raw scores to interval scores, creating a known distance between items and improving the ability to hierarchically order items based on level of ability on an underlying trait such as function and/or disability (Buz & Cortes-Rodriguez, 2016). By identifying the order of ability of items in a scale, researchers can identify gaps in a particular scale's measure of a construct (e.g., functional limitations) (Fieo, Austin, Starr, & Deary, 2011). Rasch analysis is an IRT approach that calculates the level of difficulty of individual items, and then compares item difficulty to the sample in which items were tested. This methodology provides researchers with valuable information on item and scale performance by identifying redundancy of items as well as ceiling and floor effects.

The objective of this research study was to use contemporary measurement methods to improve upon current methodologies for characterizing activity limitations within large national surveys. Our aim was to expand the breadth of activities related to overall activity limitations to include a range functional activities such as mobility, activities of daily living (ADLs), and instrumental activities of daily living (IADLs). To accomplish this aim, we developed an interval level scale of overall activity

limitation within the National Health and Aging Trends Study (NHATS). Specifically, this manuscript describes: (a) development of the NHATS-Activity Limitations scale that includes mobility, ADLs, & IADLs using IRT methods and (b) use of the newly constructed measure to characterize change in individual late-life functional limitations from 2011 to 2015.

Methods

Study Population and Data Collection

The National Health and Aging Trends Study (NHATS) is a panel study that began in 2011 which characterizes late-life disability and function among a national sample of older adults in the United States (Freedman et al., 2011; Kasper & Freedman, 2017). The NHATS is a representative sample of older adults aged 65 years and older living in the contiguous United States. The NHATS achieves a nationally representative sample through the use of complex survey design, using the Medicare Beneficiary list as the sampling frame (Freedman, 2009; Kasper & Freedman, 2017).

The NHATS uses an approach grounded in the World Health Organization's International Classification of Functioning, Disability and Health framework to capture key concepts in the disablement process including physical, sensory and cognitive capacity, accommodations, the ability to carry out essential activities independently, and participation and restrictions in valued activities (World Health Organization, 2001). The NHATS provides an opportunity to advance methodologies in estimating late-life disability and functional decline among older adults by characterizing activity limitations along a broad range of domains beyond underlying physical capacity alone (Putnam, Molton, Truitt, Smith, & Jensen, 2016).

Items administered in the survey focus on characterizing late-life functioning. The NHATS is unique in its focus on late-life functioning in that it captures not only individual level ability to perform a given activity, but includes information about difficulty in performing the activity, or need for assistance from another person or device (Freedman, Agree, Cornman, Spillman, & Kasper, 2014). The NHATS also records whether or not an individual has accommodated their performance of a difficult task by reducing the duration or frequency with which the individual has been noted to perform the activity in the past.

The baseline (Round 1) NHATS data were collected in 2011 with follow-up surveys conducted annually. For this study, we focused on developing a measure of activity limitations among community-dwelling older adults who participated in the Round 1 (2011) survey. Specifically, the items focus on the respondent's ability to perform certain self-care, mobility, and instrumental activities among individuals who are community dwelling. We exclude respondents who reported living in nursing homes given that their responses may differ systematically in the way in which they responded to these activity limitations

items used to develop the scale, creating potential bias in the overall scale scores. To develop the NHATS Activity Limitations scale, we calibrated the items using the Round 1 (2011) data. We replicated the analysis in the Round 5 (2015) data to assess the model fit as well as to examine change in activity limitations scores using the newly developed measure.

NHATS-Activity Limitations Item Construction

To achieve our goal of creating a broad representation of someone's overall activity limitations, we included items from the mobility, physical capacity, household activity, medication management, and self-care domains of the NHATS survey. We aimed to capitalize on one of the innovations the NHATS offers in its focus not only on a person's self-reported performance of the given item but its inclusion of items that address reductions in performance as well as assistance from other persons or use of devices. Devices included use of utensils for feeding, mobility devices, bath/tub seat or grab bars, dressing, and toileting devices. Accommodations included reduction in frequency since past year. Assistance from others or "help" was also considered in developing the response option coding system. In our analysis, we aimed to include this full array of function ranging from independent, adaptation, and accommodation. In order to accomplish this goal, we developed derived variables that accounted for: the level of difficulty the persons had with the item, need for help from others, use of a device, or accommodation by reducing the frequency with which a given activity was performed. In order to do this, we combined the self-reported single-item questions from the NHATS survey into 18 new items representing a range of activity limitations across all five domains of interest ([Supplementary Table 1](#) for original NHATS item description). For example, the NHATS items developed for characterizing a person's ability to go outside in the past month was originally coded as: (1) if the persons went outside by themselves (allows use of device), (2) with a device, or (3) at a reduced frequency that last year (duration accommodation). The recoded ordinal scale was: 1 = "Did not go outside"; 2 = "Had help/did not do my self"; 3 = "Had help/some difficulty doing by self"; 4 = "No help, some difficulty"; 5 = "No help, no difficulty by self, but reduced frequency/less often"; 6 = "No help, no difficulty, same or increased frequency." The response scale was constructed so that higher scores indicate better overall functional abilities (i.e., fewer activity limitations). This recoding logic was used for all items to account for the unique way in which NHATS captures self-reported late-life function within these domains.

Statistical Analysis

We first performed categorical confirmatory factor analysis using Mplus program ([Muthén & Muthén, 2005](#)) to examine

if the items from the NHATS domains (Mobility, Physical Capacity, Household Activity, Medication Management, and Self-care) could be aggregated into unidimensional factor structure model representing the concept of overall activity limitations. We assessed the unidimensionality by examining the model fit indices as Comparative Fit Index (CFI) > 0.9, Tucker Lewis Index (TLI) > 0.9, and Root Mean Square Error of Approximation (RMSEA) < 0.1 ([Chen, Curran, Bollen, Kirby, & Paxton, 2008](#); [Hoyle, 1995](#); [Hu & Bentler, 1999](#)). We examined the local independence assumption by examining whether all the absolute residual correlation values were less than 0.2 ([Reeve et al., 2007](#)). The Rasch model was used to calibrate the items and examine the item fit. Measures constructed using Rasch analysis must meet criteria of unidimensionality and have predictable hierarchies of item calibrations across the range of difficulty within the domain of assessment ([Bond & Fox, 2013](#)). Fit statistics indicated the extent to which items were fit the Rasch model, closeness of observed item scores to predicted item scores was expressed by outfit/infit mean squares (MNSQ) statistics. Both fit statistics were expected to approach 1.0, with acceptable values between 0.6 and 1.4 ([Bond & Fox, 2013](#)). Second, we estimated both item difficulty and person level function in order to examine the relative distribution of activity limitations items across range of high to low scores and assess if the newly developed scale was functioning properly. Sample score reliability was directly computed from the measurement error by using the person ability and the adjusted standard deviation (SD) of the person ability distribution ([Linacre, 2018](#)). To identify Rasch model using the Winsteps program, the average logit score of the item difficulty parameters was set as 0, and the higher scores indicate better activity limitations, representing a true interval level measure of activity limitations.

As part of the measurement development process, we aimed to ensure that the items used in the newly developed scale were acceptable from the perspective of measurement invariance. A way in which to examine this measurement property quantitatively is to examine the extent of differential item functioning (DIF) of the items. DIF captures the extent to which persons in different demographic groups (or other clinically meaningful groups) with same level of ability on a construct like function might respond differently to individual items suggesting their response is driven by more than the underlying concept itself. Estimating DIF in existing as well as newly developed measures helps support valid interpretation of group differences as either a "real" difference or an artifact of aspects of the measurement process (i.e., differing interpretation of the item by members of the different groups) ([Teresi & Fleishman, 2007](#)). To test if the items representing the Activity Limitation construct has the same conceptual meaning across NHATS subgroups, we examined DIF by U.S. region, sex, race (white vs non-white), ethnicity (non-Hispanic vs Hispanic), and age group (65–74, 75–84, and 85+). We first looked at item-level DIF

to assess statistical significance of the difference between the item difficulty parameters from different subgroups then examined the magnitude of the DIF. This multistep processes is important because DIF calculations are influenced by sample size, many items may show significant DIF for at least one comparison, even after adjustment for multiple group comparisons (Teresi, Ramirez, Lai, & Silver, 2008). Therefore, statistically significant DIF may not be clinically meaningful. This study used analysis of variance using Rasch logits to detect DIF. This method is an extension of the *t* test used to examine differences in the difficulty parameters between the groups. The following criteria were applied to determine the magnitude of DIF: items with absolute difference in item difficulties between groups ≥ 0.64 and having less than 0.05 probability to have a difference less at 0.43 signified “moderate to large” DIF; items with absolute difference in item difficulties between groups ≥ 0.43 and having less than 0.05 probability a difference at 0 indicated “slight to moderate” DIF; items with absolute difference in item difficulties between groups < 0.43 indicated “negligible” DIF (Wang, 2004). Next score-level DIF was examined to see if accounting for the item-level DIF had a significant effect on the score results using the NHATS Activity Limitations scale by generating the test characteristic curves of the Rasch models accounting for the DIF or not, and calculated the difference in the expected scores.

Lastly, to evaluate if there was change in overall activity limitations using the newly developed measure change, scores were compared from individuals between 2011 and 2015. Correlations were examined between the item difficulty parameters estimated from 2011 and those values estimated from 2015 to assess the item parameter invariance across years. Then, the person score descriptive statistics were calculated at two time points and generated the 90% confidence of minimal detectable change (MDC90), [$MDC90 = 1.65 * \text{sqrt}(2) * \text{mean of the score standard errors at 2011}$]. The MDC 90 is important in that it represents the amount of change in the score that needs to occur to be able to say that the score change exceeds expected errors in measurement with 90% confidence; thus indicating a real change in activity limitations for individual subjects (de Vet et al., 2006). A total of 3,550 subjects with complete data from two time points were used to test the comparison of change in activity limitations.

Results

Sample Characteristics

A total of 7,609 community-dwelling older adults completed the NHATS survey in the baseline 2011 cohort. Table 1 summarizes the baseline characteristics of the subjects included in the calibration of the new NHATS Activity Limitations measure.

Table 1. Calibration Sample Characteristics (*N* = 7,609)

Characteristics	<i>N</i>	%
Gender		
Male	3,171	41.67
Female	4,438	58.33
Region		
Northeast	1,403	18.44
Midwest	1,767	23.22
South	2,962	38.93
West	1,477	19.41
Race		
White	5,186	68.16
Black	1,662	21.84
Others	673	8.84
Missing	88	1.16
Ethnicity		
Non-Hispanic	7,067	92.88
Hispanic	454	5.97
Missing	88	1.16
Age		
65–74	2,988	39.27
75–84	3,018	39.66
85+	1,603	21.07

Factor Analysis and Rasch IRT Results

Based upon content coverage and parsimony, we proceeded with the one-factor solution which was inclusive of the relevant activity limitation content range as well as demonstrated acceptable fit with the data. The mean, *SD*, and the range of the item-total correlation of the 18 items were 0.69, 0.08, and 0.5–0.81. The internal consistent reliability (Cronbach coefficient alpha) equaled to 0.95. The fit indices of the unidimensional confirmatory factor model indicated the unidimensional model fit the data (CFI = 0.97, TLI = 0.966, RMSEA = 0.086). Having a low residual correlation is important to ensure validity of the construct being measured. Satisfying the local independence assumption means conditional on the latent construct being measured, the items should not be correlated with each other (or residual correlations equal to 0). Violation of this assumption could indicate an unmeasured factor that possibly accounting for the correlation between items after the impact of the target latent construct is removed. Implications of this could be the model is mis-specified potentially biasing the item parameter estimates, which in turn would bias the person score estimates. The average residual correlation for the NHATS Activity Limitation scale was -0.02 (*SD* 0.059), all the residual correlations were less than 0.2 (the range of the residual correlations are from -0.14 to 0.156); so, no item pairs violated local independence assumptions.

Five item's response categories were collapsed due to low sample size for a given response, disorder of the mean person score across some of the categories, or poor fit.

Items asking about dressing, gripping, and eating resulted in the two lowest function categories being merged: (1) “Did not do in last month due to health reason” merged with (2) “Did not by self in past month and needed help” to represent a new lowest function response category; medication management and banking resulted in the second and third categories being merged: (1) “Did by self last month with difficulty” merged with (2) “Did by self with a reduced frequency” to represent an intermediate level of reduced ability for these items. Analysis of goodness of fit of the model of the 18 NHATS items yielded infit and outfit statistics all within acceptable range for the items (Table 2). Additionally, the overall sample score reliability was good (0.83). A person-item map (Figure 1) shows graphically the hierarchy and spread of participants and items along the common linear logits across the scale range. The most difficult and easy endorsing items were “bending” and “eating.” The range of the item difficulty parameters was from -2.24 logit to 1.55 logit and the *SD* of the item difficulty parameters was 1.11 . The average person score was 1.77 logit ($SD = 1.61$), which was 1.1 person score *SD* higher than the average item difficulty. The item-person map for total scores demonstrated the distribution of items covering the lower-end of the difficulty range well with the items confirming a difficulty hierarchy consistent with expected functional limitations levels across the score range. However, the figure illustrates that the content coverage in the upper-range of score level is limited. Our analysis indicates that the NHATS activity limitations items represent tasks that are fairly easy to perform, revealing a possible ceiling effect (11% subjects had ceiling effect). This finding can also be seen

in Figure 2, which shows the test information function compared to the sample score distribution. Additionally, Figure 2 indicates that the score reliability, based on the Rasch model, for the measure is >0.90 within score range between -2.28 logit to 2.21 logits. Overall, this indicates good test score reliability for approximately 60% of the sample with gradually reducing as scores reach the upper end of the possible score range.

DIF Results

Results of DIF analyses showed significant gender DIF. No significant DIF was found for the other characteristics tested (U.S. region, race [white vs non-white], ethnicity [non-Hispanic vs Hispanic], and age). Significant gender DIF at the item level was seen for three items (laundry, carrying object, and gripping) showed “moderate to large” DIF and 1 item (bending) showed “slight to moderate” DIF. Our approach for handling any DIF found was to retain the items, rather than item deletion, and calibrate these items separately by gender to account for the underlying item-level DIF. In general, at the same logit score level, males overestimated the expected summed score than females did—results indicated the maximum expected summed score difference was at the 0.2 logit level, but the value was 1 point (compared with the summed score range from 18 to 69). This provides empirical evidence that observed gender differences for these items may reflect something other than the latent variable (i.e., bending, laundry, carrying objects, gripping), such as different interpretations of the item between males and female respondents. Based upon these results, we then evaluated the extent to which the item-level DIF had a significant impact

Table 2. Item Content, Difficulty Parameter, *SE*, and Fit Statistics

Item content	Difficulty	<i>SE</i> difficulty parameter	INFIT MNSQ	OUTFIT MNSQ
Bend	1.55	0.02	0.98	0.97
Walk	1.32	0.02	0.82	0.76
Shopping	0.9	0.02	0.9	0.83
Carrying	0.89	0.02	0.78	0.66
Stairs	0.89	0.02	0.83	0.65
Gripping	0.75	0.03	1.11	1.2
Meal prep	0.69	0.02	1.02	0.98
Laundry	0.68	0.02	1.08	1.12
Banking	0.39	0.02	1.07	1.15
Reaching	0.29	0.02	1.03	0.96
Medication	-0.1	0.02	1.11	1.25
Bathing	-0.25	0.02	1.05	0.88
Dressing	-0.35	0.02	1.31	1.34
Outdoor mobility	-0.47	0.02	0.98	0.86
Bed mobility	-1.39	0.03	0.91	0.85
Indoor mobility	-1.66	0.02	1.13	0.97
Toileting	-1.89	0.03	0.84	0.66
Eating/feeding	-2.24	0.03	1.17	1.31

Figure 1. Rasch item map of sample and item distribution. Rasch item map description: Column labeled "Sample" shows subject distribution across the scale. Each "#" represents 73 subjects. Column labeled items displays the item distribution by level of difficulty. "M" is mean for items and sample. "S" is one SD of each distribution. Each item content label is listed along with its average level of difficulty.

on the overall scale score by examining the test characteristic curves based on whether the item parameters adjusted the DIF or not, and examined the difference in the curves. We did not find substantive magnitude DIF at the overall score level; thus, we did not consider DIF in following analyses comparing overall changes in NHATS Activity Limitation scores over time.

NHATS Activity Limitations Change 2011 Versus 2015

Using the newly created metric of activity limitations, we sought to examine change over time by comparing individual scores among individuals who completed the NHATS survey in 2011 and 2015. The NHATS cohort of individuals who completed the items composing the activity limitations measure in both 2011 and 2015 was 3,550. The scatter plot of item difficulty parameters (Figure 3) based on 2011 and 2015 illustrates that the item parameter estimates based on 2011 were highly correlated with the values in based on 2015 and the values were distributed around $y = x$ line. This finding supports the assumption that

Figure 2. Test information function, reliability, and sample score distribution. The x-axis presents the Logit scores (0 means the average item difficulty equals to 0). The unit of the left y-axis is the value of the Test Information Function (TIF) (where the function scale provides the most information along the score distribution.). The unit of the right y-axis is the value of the sample proportion. The bell shape histogram is the TIF. The black histogram is the sample score distribution in the 2011 National Health and Aging Trends Study (NHATS) sample. The dash line indicates the test information value which corresponds to the score reliability = 0.9.

Figure 3. Correlation of item difficulty parameters for National Health and Aging Trends Study (NHATS) Activity Limitation Measure (2011 and 2015).

item difficulties are invariant across years 2011 and 2015. Item difficulty invariance across test administrations (i.e., 2011 vs 2015) is important to ensure that any resultant change score is a function of possible "real" change versus item measurement properties. The average person score measurement error was 0.59 in 2011 with the calculated MDC90 of 1.38. This can be interpreted in that there is 90% confidence that a 1.38 change in the NHATS Activity Limitations score represents "real" change versus a change due to measurement error. Using this MDC90 criteria, 3.46% of the sample had a change in score $>$ MDC90 (i.e.,

improve in function/decreased activity limitations); 15.83% had a negative change (i.e., decrease in function/increased activity limitations); and 80.70% had no change exceeding $\pm$ MDC 90 (i.e., no change in function/activity limitations).

Discussion

Results of this study support the development of a new measure of activity limitations within the NHATS data. By using a Rasch measurement development approach, we confirmed that activity limitations, capturing aspects of mobility, physical capacity, ADLs, and IADLs, can be measured on a unidimensional interval scale. The resulting items in the scale were ordered in the expected level of difficulty and were consistent with both theoretical and clinical understandings of functional activity limitations (i.e., tasks like bed mobility are at the lower or “easier” end of the scale while items such as meal preparation are at the higher end of the difficulty scale). This finding supports initial construct validity of the NHATS Activity Limitations measure we developed in this study. The NHATS Activity Limitation scale calibrates activity limitations of 18 items (functional mobility, ADLs, IADL) along the same underlying metric which incorporates variation in item difficulty into the score, rather than assuming each item is equally difficult. This is a significant methodological advancement to common approaches used in national surveys of self-reported older adult disability and health.

In terms of content coverage, this measure represents a broader range of activity limitations items than has previously been developed. Additionally, we capitalized on the unique data collection framework used in NHATS to capture not only limitations in self-reported activity performance but also modifications (using help or with assistive devices) as well as accommodation (reducing the frequency of performing a given task). Assistive device use and accommodations allowed a finer gradation of response categories to be included in the scaled score. While both decrease in frequency of activity performance (accommodating) and device use within NHATS items represents an advancement over current survey methods, additional exploration of device use and accommodations within context of the activity being performed may prove useful in detecting functional decline at the higher upper end of the score ranges (among higher functioning older adults). Relatedly, study results revealed ceiling effects for the newly developed scale. [Figure 2](#) illustrates that the majority of the NHATS items represent “easier” tasks while a large proportion of the sample is functioning at the higher end of the scale (not having difficulty with these easier activities). This finding indicates that future item development targeting higher functioning individuals may be useful to increase the scale’s sensitivity to change and possibly allow detection of preclinical disability levels among community-dwelling older adults. For example,

adding items that increase the intensity or duration of the activity performance such as “heavy household cleaning, home maintenance and repair” are examples more difficult activities that may reduce the current ceiling effect observed in the scale. Apart from problems with ceiling effects, which have been found in previous studies as well ([Kasper et al., 2017](#)), the items represent a range of content coverage for characterizing overall functional limitations among community-dwelling older adults across domains including mobility, physical capacity, ADLs, and IADLs.

The Rasch IRT-based measure allows a composite score to represent a broad range of ability along a true interval scale. This is an advantage over typical measures of functional activity limitations that have historically been used in national surveys where the measures are limited to estimating only individual item change or creating an arbitrary summed scale which assumes that each item is as difficult as any other given item. Furthermore, we found that some items demonstrate significant DIF based on gender (male vs female). While the results of significant DIF canceled at the aggregate measure level, this finding may have important implications at the item, individual level. This finding is consistent with previous literature highlighting the importance of considering the impact of DIF at both the item level and overall score level ([Crane, Gibbons, Jolley, & van Belle, 2006](#)).

Using the NHATS Activity Limitations measure, we found that a small proportion of these individuals showed significant improvements in functional ability level from 2011 to 2015, a larger proportion showed a significant decline in functional activity ability, while the majority of individuals showed no significant change. Although the change is modest, using the Rasch-based model to examine change in a more precise way allows identification of whether changes in score are of comparable importance according to the underlying construct by accounting for variation in item difficulty across the range of the possible scale scores. In a Likert scale model, no strict item hierarchy is hypothesized or defined and priority is given to internal consistency of the items within the scale ([Stucki et al., 1996](#)).

Limitations

Within NHATS, we see clear limitations in our ability to characterize activity limitations in the U.S. population aged 65+ at the higher levels of functional ability. Given the large ceiling effect, there may be more decline in activity limitations at the upper end of the scale where the scores have larger standard errors and our ability to characterize limitations is limited. One way to improve upon the NHATS Activity Limitations measure would be to add items at the upper end of the difficult range to the NHATS survey to address the ceiling effects. Measures developed using IRT-based methods are designed to easily add items and replenish the item banks to improve both

content coverage and psychometric properties of a measure (Marfeo et al., 2017; McDonough et al., 2017). Such future work would be appropriate for improving the ability to characterize late-life function among a national sample of older adults within the existing NHATS infrastructure. Specifically, an optimal way to improve the scale would be to add new items to fill in gaps along the scale and to delete redundant items representing the same level of difficulty. This future work would allow investigation of the extent to which the large proportion of individuals staying the same (no change in activity limitation) was due to lack of scale reliability at the upper score ranges or true lack of sensitivity to change.

Implications

This study demonstrates that by using IRT methods a unidimensional, interval scale of self-reported late-life functional activity limitations can be constructed using traditional survey measures nested within the National Health and Aging Trends Study. The resulting NHATS-Activity Limitation scale demonstrated acceptable scaling properties. Implications of this study include (a) the use of the NHATS Activity Limitations scale as an IRT-based scale for characterizing the magnitude and direction of perceived functional change in late-life activity limitations and (b) show promise for improving current estimates of late-life disability and functional decline. Results did reveal concerns regarding ceiling effects within the current NHATS self-report items, suggesting future work is needed to expand the range of ability currently represented in the NHATS activity limitation items.

Supplementary Data

Supplementary data are available at *The Gerontologist* online.

Funding

This study was supported by National Institutes of Health – National Center for Medical Rehabilitation Research (NICHD) and the National Institute of Neurological Disorders and Stroke through the Center for Large Data Research and Data Sharing in Rehabilitation (CLDR) pilot studies program (Grant/Award Number: NIH R24 HD065702).

Conflict of Interest

None reported.

Ethical Review

This study was exempt from review due to not using human subjects in this research project. All data used are deidentified and publicly available via www.nhats.org.

References

- Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Psychology Press.
- Buz, J., & Cortes-Rodriguez, M. (2016). Measurement of the severity of disability in community-dwelling adults and older adults: Interval-level measures for accurate comparisons in large survey data sets. *BMJ Open*, 6, e011842. doi:10.1136/bmjopen-2016-011842
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36, 462–494. doi:10.1177/0049124108314720
- Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, 44, S115–S123. doi:10.1097/01.mlr.0000245183.28384.ed
- Fico, R. A., Austin, E. J., Starr, J. M., & Deary, I. J. (2011). Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: A systematic review. *BMC Geriatrics*, 11, 42. doi:10.1186/1471-2318-11-42
- Freedman, V. A. (2009). Adopting the ICF language for studying late-life disability: A field of dreams? *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences*, 64, 1172–1174; discussion 1175. doi:10.1093/gerona/glp095
- Freedman, V. A., Agree, E. M., Cornman, J. C., Spillman, B. C., & Kasper, J. D. (2014). Reliability and validity of self-care and mobility accommodations measures in the National Health and Aging Trends Study. *The Gerontologist*, 54, 944–951. doi:10.1093/geront/gnt104
- Freedman, V. A., Kasper, J. D., Cornman, J. C., Agree, E. M., Bandeen-Roche, K., Mor, V., ... Wolf, D. A. (2011). Validation of new measures of disability and functioning in the National Health and Aging Trends Study. *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences*, 66, 1013–1021. doi:10.1093/gerona/glr087
- Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications, Inc.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Kasper, J. D., Chan, K. S., & Freedman, V. A. (2017). Measuring physical capacity. *Journal of Aging and Health*, 29, 289–309. doi:10.1177/0898264316635566
- Kasper, J. D., & Freedman, V. A. (2017). *National health and aging trends study user guide: Rounds 1–6 final release*. Baltimore, MD: Johns Hopkins University School of Public Health.
- Linacre, J. M. (2018). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.
- Marfeo, E. E., Ni, P., McDonough, C., Peterik, K., Marino, M., Meterko, M., ... Jette, A. M. (2017). Improving assessment of work related mental health function using the work disability functional assessment battery (WD-FAB). *Journal of Occupational Rehabilitation*, 28, 190–199. doi:10.1007/s10926-017-9710-5.

- McDonough, C. M., Ni, P., Peterik, K., Marfeo, E. E., Marino, M. E., Meterko, M.,...Chan, L. (2017). Improving measures of work-related physical functioning. *Quality of Life Research*, *26*, 789–798. doi:10.1007/s11136-016-1477-1
- Muthén, L. K., & Muthén, B. O. (2005). *Mplus: Statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.
- Nagi, S. Z. (1976). An epidemiology of disability among adults in the United States. *The Milbank Memorial Fund quarterly. Health and Society*, *54*, 439–467. doi:10.2307/3349677
- das Nair, R., Moreton, B. J., & Lincoln, N. B. (2011). Rasch analysis of the Nottingham extended activities of daily living scale. *Journal of Rehabilitation Medicine*, *43*, 944–950. doi:10.2340/16501977-0858
- Putnam, M., Molton, I. R., Truitt, A. R., Smith, A. E., & Jensen, M. P. (2016). Measures of aging with disability in US secondary data sets: Results of a scoping review. *Disability and Health Journal*, *9*, 5–10. doi:10.1016/j.dhjo.2015.07.002
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A.,...Cella, D.; PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*, S22–S31. doi:10.1097/01.mlr.0000250483.85507.04
- Resnick, B., Galik, E., Dorsey, S., Scheve, A., & Gutkin, S. (2011). Reliability and validity testing of the physical resilience measure. *The Gerontologist*, *51*, 643–652. doi:10.1093/geront/gnr016
- Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press. doi:10.1093/med/9780199685219.001.0001
- Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*, *49*, 711–717. doi:10.1016/0895-4356(96)00016-9
- Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16* (Suppl 1), 33–42. doi:10.1007/s11136-007-9184-6
- Teresi, J. A., Ramirez, M., Lai, J. S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly*, *50*, 538.
- de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, *4*, 54. doi:10.1186/1477-7525-4-54
- Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*, 221–261.
- Wijers, I. G., Ayala, A., Rodriguez-Blazquez, C., Rodriguez-Laso, A., Rodriguez-Rodriguez, V., & Forjaz, M. J. (2017). Rasch analysis and construct validity of the disease burden morbidity assessment in older adults. *The Gerontologist*, *58*, e302–e310. doi:10.1093/geront/gnx061
- World Health Organization. (2001). *International classification of functioning, disability and health: ICF*. Geneva, Switzerland: World Health Organization.