The Nursing Home Physical Performance Test: A Secondary Data Analysis of Women in Long-Term Care Using Item Response Theory

Subashan Perera, PhD, David A Nace, MD, Neil M Resnick, MD, Susan L Greenspan, MD
The Nursing Home Physical Performance Test: A Secondary Data Analysis of Women in Long-Term Care Using Item Response Theory

Subashan Perera, PhD, David A. Nace, MD, Neil M. Resnick, MD, and Susan L. Greenspan, MD

Division of Geriatric Medicine and Department of Biostatistics, University of Pittsburgh, Pennsylvania.

*Address correspondence to Subashan Perera, PhD, Division of Geriatric Medicine, University of Pittsburgh, 3471 Fifth Avenue, Suite 500, Pittsburgh, PA 15213. E-mail: ksp9@pitt.edu

Received: January 31, 2017; Editorial Decision Date: March 3, 2017

Decision Editor: Rachel Pruchno, PhD

Abstract

Background and Objectives: The Nursing Home Physical Performance Test (NHPPT) was developed to measure function among nursing home residents using sit-to-stand, scooping applesauce, face washing, dialing phone, putting on sweater, and ambulating tasks. Using item response theory, we explore its measurement characteristics at item level and opportunities for improvements.

Research Design and Methods: We used data from long-term care women. We fitted a graded response model, estimated parameters, and constructed probability and information curves. We identified items to be targeted toward lower and higher functioning persons to increase the range of abilities to which the instrument is applicable. We revised the scoring by making sit-to-stand and sweater items harder and dialing phone easier. We examined changes to concurrent validity with activities of daily living (ADL), frailty, and cognitive function.

Results: Participants were 86 years old, had more than three comorbidities, and a NHPPT of 19.4. All items had high discrimination and were targeted toward the lower middle range of performance continuum. After revision, sit-to-stand and sweater items demonstrated greater discrimination among the higher functioning and/or greater spread of thresholds for response categories. The overall test showed discrimination over a wider range of individuals. Concurrent validity correlation improved from 0.60 to 0.68 for instrumental ADL and explained variability ($R^2$) from 22% to 36% for frailty.

Discussion and Implications: NHPPT has good measurement characteristics at the item level. NHPPT can be improved, implemented in computerized adaptive testing, and combined with self-report for greater utility, but a definitive study is needed.

Keywords: Function/mobility, Long-term care, Measurement, Psychometrics
et al., 2009; Oh-Park, Perera, & Verghese, 2012; Perera, Mody, Woodman, & Studenski, 2006; Perera et al., 2014). Physical performance measures are increasingly being used as outcomes in randomized trials of the community-dwelling older adults (Brach, VanSwearingen, Perera, Wert, & Studenski, 2013; Pahor et al., 2014; VanSwearingen et al., 2009).

In contrast, the evidence base for physical performance assessment in frail long-term care older adults is less developed due to many challenges. For example, assessments based on walking or climbing stairs that are frequently used in community-dwelling older adults may be too difficult and unsafe for long-term care residents. Nursing Home Physical Performance Test (NHPPT) is one measure of physical performance developed specifically for long-term care residents and seeks to address some of the challenges (Binder, Miller, & Ball, 2001). On an ordinal scale of 0–4 (higher score better), it assesses level of assistance and prompting required as well as time required (quantitative scoring) to complete sit-to-stand, scooping applesauce, face washing, dialing a telephone, putting on and taking off a sweater, and ambulating 6 meters (Supplemental Appendix Table 1). The tasks are more relevant for the long-term care older adults, and the ambulation task is allowed to be completed with a wheelchair if unable to walk. NHPPT has been used as a secondary outcome in a randomized trial of nursing home and assisted living residents examining effects of activities and exercise on sleep (Lorenz et al., 2012; Richards et al., 2011). More recently, NHPPT was included as a secondary outcome in a trial of zoledronic acid to improve bone mineral density (BMD) in a similar population (Greenspan, Perera, Ferchak, Nace, & Resnick, 2015).

The NHPPT was developed, and its psychometric properties such as reliability, validity, and internal consistency have been evaluated using classical methods (Binder et al., 2001). NHPPT has excellent interrater reliability (0.73–0.93); consists of underlying factors of gross motor function as well as fine motor coordination and task sequencing; high internal consistency; has concurrent validity with respect to other available measures of function and frailty.

Research Design and Methods

Setting and Participants
In the present secondary analysis, we used cross-sectional data from women that were screened for participating in a randomized trial of zoledronic acid to improve BMD. The methods of the parent study have been previously described in detail elsewhere (Greenspan et al., 2012, 2015). Briefly, participants were recruited to the parent trial from nursing homes and assisted living facilities. Inclusion criteria included not receiving a bisphosphonate, history of vertebral or hip fracture, a BMD lower than 2003 National Osteoporosis Guidelines for treatment, and a BMD lower than 2 SD from a healthy 30-year-old adult (National Osteoporosis Foundation, 2003). Those with cognitive and functional impairment, immobility, multimorbidity, and polypharmacy were not excluded. Only those with a life expectancy of <2 years and estimated glomerular filtration ratio of <30 mL/min were excluded. We included 205 participants with NHPPT data in the present analysis.

Measures
We considered the baseline measurements of quantitatively assessed NHPPT (Binder et al., 2001), Katz Activities of Daily Living Scale (ADL; Katz et al., 1963), Instrumental Activities of Daily Living (IADL; Lawton & Brody, 1969), Short Portable Mental Status Questionnaire (SPMSQ; Pfeiffer, 1975), Comorbidity Index (Rigler, Studenski, Wallace, Reker, & Duncan, 2002), Patient Health Questionnaire for depression (PHQ-9; Kroenke & Spitzer, 2002), and in a subset of 176 participants, the modified Fried Frailty Index (Fried et al., 2001) categorized as robust, prefrail, and frail.

Statistical Analysis
We used appropriate descriptive statistics to summarize the sample characteristics. First, we performed an item response analysis of the NHPPT data using a
graded response model, suitable for item analysis when responses are ordinal (De Ayala, 2009; Samejima, 1969). The basic model developed for dichotomous (correct/incorrect) responses (i.e., the one-parameter Rasch model) stipulates the probability of a correct response to a question depends on both a person’s ability and the question’s difficulty, measured on the same underlying latent trait the test is designed to measure (Rasch, 1980). It has been extended to the graded response model for use with ordinal responses, where additional parameters have been added to accommodate different levels of a “correct” response (thresholds), and the level of discrimination afforded by an item. For the NHPPT scale where each item has an ordinal score from 0 to 4, we used the graded response model given by

\[ P_{ijk}^* = \frac{e^{\lambda_j \eta_i - \alpha_{jk}}}{1 + e^{\lambda_j \eta_i - \alpha_{jk}}} \],

where \( P_{ijk}^* \) is the probability of participant \( i \) responding at level \( k \) or greater for item \( j \); \( \alpha_{jk} \) are the threshold parameters for item \( j \); \( \lambda_j \) is the slope or discrimination parameter for item \( j \); \( \eta_i \) is the normally distributed latent trait of the \( i \)th participant at NHPPT attempts to measure. Item category probability \( P_{ijk} \), the probability of participant \( i \) responding at level \( k \) for item \( j \), can then be estimated by taking differences \( P_{ijk}^* - P_{ij(k-1)}^* \) for adjacent categories, where \( P_{ij0}^* = 0 \) and \( P_{ij4}^* = 1 \). Parameters were estimated; item category probability, item information, and test

**Table 1. Participant Characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Mean ± SD or N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>85.8 ± 5.2</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Elementary school</td>
<td>7 (3.4%)</td>
</tr>
<tr>
<td>High-school equivalent</td>
<td>106 (51.7%)</td>
</tr>
<tr>
<td>College</td>
<td>51 (24.9%)</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>25 (12.2%)</td>
</tr>
<tr>
<td>Other</td>
<td>14 (6.8%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (1.0%)</td>
</tr>
<tr>
<td>ADL score (range 0–14; higher is better)</td>
<td>11.7 ± 2.3</td>
</tr>
<tr>
<td>IADL score (range 0–14; higher is better)</td>
<td>8.0 ± 3.8</td>
</tr>
<tr>
<td>SPMSQ errors (range 0–10; higher is worse)</td>
<td>1.9 ± 2.4</td>
</tr>
<tr>
<td>PHQ-9 (range 0–27; higher is worse)</td>
<td>3.8 ± 4.2</td>
</tr>
<tr>
<td>Comorbidity index (range 0–8; higher is worse)</td>
<td>3.5 ± 1.4</td>
</tr>
</tbody>
</table>

**Table 2. Graded Response Model Parameter Estimates ± Standard Error**

<table>
<thead>
<tr>
<th>Item</th>
<th>Original scoring</th>
<th>Revised scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope/discrimination parameter (( \lambda_j ))</td>
<td>Threshold parameters (( \alpha_{jk} ))</td>
</tr>
<tr>
<td>1. Sit-to-stand</td>
<td>2.20 ± 0.38</td>
<td>0: −2.09 ± 0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −2.01 ± 0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −1.40 ± 0.18</td>
</tr>
<tr>
<td>2. Scooping applesauce</td>
<td>1.73 ± 0.32</td>
<td>0: −2.97 ± 0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −2.46 ± 0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −1.98 ± 0.28</td>
</tr>
<tr>
<td>3. Face washing</td>
<td>1.37 ± 0.28</td>
<td>0: −3.81 ± 0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −2.92 ± 0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −1.71 ± 0.29</td>
</tr>
<tr>
<td>4. Dial a telephone</td>
<td>1.34 ± 0.26</td>
<td>0: −2.38 ± 0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −2.18 ± 0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −1.31 ± 0.23</td>
</tr>
<tr>
<td>5. Put on/take off sweater</td>
<td>2.19 ± 0.47</td>
<td>0: −1.67 ± 0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −1.63 ± 0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −1.42 ± 0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3: −1.07 ± 0.16</td>
</tr>
<tr>
<td>6. 6-meter walk/wheel</td>
<td>2.21 ± 0.39</td>
<td>0: −1.79 ± 0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: −1.09 ± 0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: −0.07 ± 0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3: 0.77 ± 0.13</td>
</tr>
</tbody>
</table>
Information curves were plotted against the estimated latent trait; and a histogram of estimated latent trait was made. Second, depending on the observations from the parameter estimates and item curves, and substantive judgment, the published thresholds for quantitative scoring were tweaked with the goals of targeting some items toward frailty (making easier) and some toward higher functioning (making more difficult) and increasing item information for participants with a wider range of function. Third, we fitted the graded response model to the revised NHPPT. Finally, we examined concurrent validity of both original and revised scoring of NHPPT with respect to frailty and other measures of function and mood using correlation and regression analyses. SAS® IRT procedure was used for the main analysis (SAS Institute, Cary, NC).

Results

Participant Characteristics
Participants had a mean age of 86 years, and the majority had had a high-school education (Table 1). Overall, participants had high levels of functional impairment. Mean NHPPT score was 19.4, and on average, participants had more than three comorbidities. Mean ADL and IADL scores were 11.7 and 8.0, respectively.

Item Characteristics of the NHPPT
Table 2 shows that the slope parameters are generally high for all items, ranging from 1.34 for Dial a Telephone to 2.21 for 6-meter walk/wheel, indicating that probability of endorsing an item changes rapidly with a person’s ability thus affording a greater level of discrimination. In addition, threshold parameters for all items appear to have a generally wide range from high negative values to either low negative values or somewhat positive values, indicating that items can discriminate among those with a wider range of person abilities toward the frail end of the spectrum of physical performance. The narrowest of the range was for Put on/Take off Sweater, but still ranged from −1.67 to −1.07. Item category probabilities for each item are plotted against person ability in Figure 1. Different categories (0/1/2/3/4) are largely endorsed with substantial probability by persons with some range of abilities, as exemplified in the Walk/Wheel item. But for some items such as Sit-to-Stand, Dial a Telephone, and Put on/Take off Sweater, no level of ability elicits a substantial probability of performing at the Category 1 level. The extent of information provided by each item, to be interpreted on a comparative basis rather than absolute numbers, is in Figure 2a. Most items provide greatest information for discriminating among lower end of functional performance (∼4 to 0 range), whereas Sit-to-Stand and Walk/Wheel in the moderate −2 to +2 range.

Opportunities for Revising NHPPT Scoring
Scoop Applesauce and Washing Face items are most relevant and geared toward those at low end of the performance continuum, whereas Walk/Wheel is most applicable to those in the middle ranges. Figure 2a shows these items already have somewhat wide information curves. Therefore, these items probably do not represent much opportunity for improvement. In a classroom test, it is desirable to have items that have a wide range of difficulty. Similarly, a good instrument should have items with a wide range in terms of difficulty/ease to endorse. Sit-to-Stand and Put on/Take off Sweater items appear candidates for revising their scoring so that they are harder to endorse, and Dial a Telephone easier to endorse. For example, we changed the quantitative scoring so that Sit-to-Stand task should be completed within 1 s rather than 2 s to receive a score of 4, and Dial a Telephone could be completed within 20 s rather than 13 s to receive a score of 4 (Supplemental Appendix Table 1). After making the said changes to item difficulties by changing quantitative scoring, the threshold parameters are more spread out (Table 2). Figure 3b shows that Sit-to-Stand and Put on/Take off Sweater items now provide a greater level of information for discriminating among more high functioning individuals of the performance continuum in the −1 range.
to +3 range. *Dial a Telephone* item still targets the same range of individuals but provides much greater information. As a complete test, NHPPT with revised scoring provides a greater level of information for individuals with a wider range of abilities. For example, if we arbitrarily consider an amount of information greater than 3, the range of corresponding abilities have widened from approximately (−3 to +1) to (−3 to +2), substantially extending the utility of the test among the higher functioning individuals. Finally, histograms of participants’ estimated latent traits show that some of the ceiling effects present in the original scoring have been alleviated by revised scoring, resulting in a desirable symmetric shape consistent with the normality assumption of participant abilities $\eta_i$ in the model (Figure 4).

**Concurrent Validity**

With the revision of scoring, the correlation of NHPPT with ADL score increased from 0.67 to 0.68, IADL score from

![Figure 2. Item information curves for original and rescored Nursing Home Physical Performance Test versus physical performance trait.](image)

![Figure 3. Test information curve for original and rescored Nursing Home Physical Performance Test versus physical performance trait.](image)

![Figure 4. Histogram of estimated latent traits from the original and revised Nursing Home Physical Performance Test.](image)
Discussion and Implications

Using methods of item response theory, we have shown that the NHPPT largely has good psychometric properties at the item level. Items have good discriminatory properties indicated by the slope parameters and are targeted mostly toward the lower end of the physical performance spectrum. NHPPT has concurrent validity with respect to other measures of physical and cognitive function, and frailty. Opportunities were identified for further improvement of the NHPPT that could be accomplished by rescoring. The score revision is shown to improve performance of the instrument at both item and overall test level and also increases the concurrent validity with respect to other measures considered.

Objectively measured and scored performance-based measures such as NHPPT are especially appealing compared with self-reports of function particularly in nursing home residents due to high prevalence of cognitive impairment, reporting bias, and recall bias and that residents simply do not engage in many of the everyday activities on which most self-report questionnaires of function are based. Item analysis of NHPPT is an important next step for several reasons. First, having actually used NHPPT in a randomized trial, we have encountered some practical challenges with some items in the targeted frail population. For example, Putting on/Taking off Sweater caused us to think about whether to use a research set of sweaters in standard sizes. Participants may be uncomfortable using sweaters worn by others but using a resident’s own sweater would introduce substantial variability. Furthermore, sweaters may be uncomfortable during warmer months of the year. Moreover, scooping applesauce can be messy and challenging for some residents depending on their eating habits and cognitive function. If a challenging item also turns out to have poor psychometric characteristics from an item analysis, perhaps it can be revised or skipped. Second, item response theory naturally lends itself to computerized adaptive testing (McDonough et al., 2012) where only a subset of items specially targeted for an individual’s ability is administered. The score derived from the shorter test generally has similar properties to the full test. Shorter and targeted physical performance tests may be especially advantageous in frail populations such as ours where fatigue and safety are important considerations. With the technological advancement new ways of functional assessment are now possible (Lowe et al., 2013), and computerized adaptive testing may play an important role in the said evolution. Third, self-report questionnaires and performance-based measures often provide complementary information about the functional status of the elderly (Perera et al., 2005). Combination of performance measure based items from NHPPT and items from other established self-report questionnaires can be examined using the methods of item response theory to potentially develop new scales that can reap advantages of both assessment paradigms.

Our study has several important strengths. First, to our knowledge, it is the first analysis of NHPPT using item response theory. In fact, application of item response theory in assessment of older long-term care residents thus far has been sparse. We are aware of only a few studies involving assessment of pain and Barthel Index (Edelen & Saliba, 2010; Liu, Unick, Galik, & Resnick, 2015; van Nispen tot Panne rerden et al., 2009). Second, it is also the largest study thus far to use NHPPT. The original development and validation work was based on only 95 participants, and a sleep and exercise trial that used NHPPT as an outcome had only 119 participants (Binder et al., 2001; Lorenz et al., 2012; Richards et al., 2011). Apart from gender, our study participants appear to be fairly comparable to those used for original scale development. Third, we were able to revise scoring in an informed manner based on guidance provided by the item analysis. We feel that it is especially appealing that high-end discrimination was improved merely by scoring rather than addition of more challenging tasks such as climbing steps, as the latter approach has the disadvantages of safety concerns and making the instrument longer than necessary. Finally, we examined test information and concurrent validity and showed substantial improvements with the revised scoring.

Our study has several important limitations as well. First, our parent study only included women. How well the results will generalize to the general long-term care population, including men, is not known and difficult to speculate. Studies in community-dwelling older adults show that men perform higher than women on physical performance measures such as gait speed and have lower rates of disability than women (Perera et al., 2016). However, such findings are also influenced by greater longevity in men and the resulting survival bias (Studenski et al., 2011). Second, an even larger sample size is needed for a more comprehensive and definitive item analysis of the NHPPT. Third, we had found that only a very small number of participants from the entire continuum was performing at level that would yield Category 1 for Sit-to-Stand, Dial a Telephone, and Put on/ Take off Sweater items. The category can likely be combined with either Category 0 or Category 2 without much loss of information. However, we did not revise this aspect of scoring in this initial investigation with a limited sample because it would have changed the range of scores for NHPPT.

We conclude that NHPPT has good psychometric properties from an item response theoretic perspective, which complements the insights obtained through classical test theory. Through our item analysis, we have identified some opportunities for improvement of NHPPT scoring and demonstrated such revisions can yield improved test information characteristics and concurrent validity. Further
study with a larger more representative sample applying item response theory is needed to make more definitive recommendations for improving NHPPPT and possibly adapting it for computerized adaptive testing and/or combining with self-report items.

Supplementary Material

Supplementary data is available at The Gerontologist online.

Funding

This work was supported by National Institutes of Health (R01 AG028068, K24 DK062895, and P30 AG024827 to S. L. Greenspan).

Conflict of Interest

S. Perera and S. L. Greenspan receive research funding through a grant from Eli Lilly & Co. to the University of Pittsburgh. S. L. Greenspan receives research funding from Amgen to the University of Pittsburgh and is on the advisory board of Merck. D. A. Nace receives research funding from Amgen to the University of Pittsburgh for influenza research.

References


for long-term care residents. *Nursing Research*, 64, 88–99. doi:10.1097/NUR.0000000000000072


