Development and Validation of a Video-Animated Tool for Assessing Mobility

W. Jack Rejeski, Edward H. Ip, Anthony P. Marsh, and Ryan T. Barnard

Background. Existing self-report measures of mobility ignore important contextual features of movement and require respondents to make complex judgments about specific tasks. Thus, we describe the development and validation of a short form (sf) video-animated tool for assessing mobility, the Mobility Assessment Tool—MAT-sf.

Methods. This study involves cross-sectional and longitudinal analyses examining the measurement properties of the MAT-sf. The MAT-sf consists of 10 animated video clips that assess respondents’ level of proficiency in performing each task. The main outcome measures used for validation included the Pepper Assessment Tool for Disability (PAT-D), the Short Physical Performance Battery (SPPB), and 400-m walk test.

Results. Participants (n = 234), 166 women and 68 men, had an average age of 81.9 years and a variety of comorbidities with 65.4% having high blood pressure. An average SPPB score of 8.6 (range 2–12) suggests that the study sample had evidence of compromised physical function but was quite heterogeneous. The MAT-sf had good content validity, excellent test–retest reliability (r = .93), and criterion-related validity with the PAT-D. Moreover, the MAT-sf added considerable variance to the prediction of both SPPB scores and 400-m gait speed over and above the PAT-D mobility subscale. The MAT-sf also discriminated between older adults who completed or failed the 400-m walk test.

Conclusion. The MAT-sf is an innovative psychometrically sound measure of mobility. It has utility in epidemiological studies, translational science, and clinical practice.

Key Words: Mobility—Aging—Measurement—Disability.

Received November 1, 2009; Accepted March 11, 2010

Decision Editor: Luigi Ferrucci, MD, PhD

The onset of difficulty with mobility marks a serious decline in quality of life (1), is central to sustaining independence with aging (2,3), and confers an increased risk of institutionalization and death (4,5). Its importance as a topic in geriatric medicine cannot be overstated. The most common measures of mobility in the research literature have been the Short Physical Performance Battery (SPPB) (6) and walking tasks, such as the 400-m walk test (5). However, because performance-based tests are not always possible, and in fact convey somewhat different information than people’s perceptions of their capacities, self-report measures have been developed, including measures such as the physical functioning subscale of the Short Form 36-item Health Survey (SF-36) (7) and the Pepper Assessment Tool for Disability (PAT-D) (8). Common limitations of existing self-report measures are that respondents are required to make complex judgments about the meaning of specific tasks, and contextual factors are ignored that are important in making task-related judgments. For example, when asked about difficulty or limitations in walking a block, how far is a block and what speed is expected? Does walking a block include hills or involve uneven terrain? The lack of specificity in item content creates error in measurement and fails to adequately capture the demands of many tasks.

The objectives of this initial phase of our research are to describe the process that was used in the selection of items for a short form (sf) Mobility Assessment Tool—the MAT-sf—that employed animated video clips to provide standardized representation of task performance along dimensions of speed, the slope of inclines, and other environmental challenges. State-of-the-art methods from item response theory (IRT) were employed to achieve this task. We then examine the reliability and validity of the MAT-sf, establishing concurrent validity, and demonstrating that the measure outperforms the PAT-D in predicting physical performance scores for the SPPB and 400-walk test. The value of the PAT-D in comparison with a measure such as the SF-36 is that there are subscales for mobility, basic
activities of daily living (ADLs), and instrument activities of daily living (IADLs); the measure also has item content that mirrors other major self-report measures of mobility because this strategy served as the basis for the development of the PAT-D.

**Methods**

**Participants**

The sample for the initial scale development included 234 men and women who lived independently in either the community or continuing care retirement communities in North and South Carolina. Participants were recruited via advertisements approved by the institutional review board (IRB), community social/wellness directors, and lectures on aging given by senior investigators on the project. The methods for this investigation were approved by our institutional IRB for human subject research. The inclusion criteria included age between 65 and 90 years, able to walk without help either with or without the use of an assistive device, Mini-Mental State Exam \( \geq 23 \), and willing to give an informed consent and sign a Health Insurance Portability and Accountability Act of 1996 authorization form. Participants were excluded if they had one or more of the following: undergoing active treatment for a psychiatric illness, severe symptomatic heart disease that might put them at risk during the 400-m walk test, resting blood pressure \( > 160/100 \) mmHg, and severe systemic diseases.

**Item Development and Selection**

We began by creating 81 video clips for mobility that encompassed a wide range of tasks characterized by demands that ranged from low to high. Our long-term goal is to develop a computerized adaptive testing protocol using all items. The original 81 items were subjected to an IRT-based calibration procedure (9), employing a two-parameter logistic model. The software program MULTILOG (10) was used to perform the calibration. Item parameters were estimated through a maximum likelihood procedure (11) enabling us to set the scale for the MAT-sf. The items were then partitioned into seven functional clusters [number of items]: cluster A [10]—ambulating with a cane as an assistive device, cluster B [5]—walking on a flat surface at different speeds, cluster C [6]—walking up inclined ramps, cluster D [2]—walking while stepping over hurdles, cluster E [6]—walking outdoors uphill on uneven terrain, cluster F [26]—climbing stairs with and without handrails of different runs, and cluster G [26]—climbing stairs while carrying bags in one or both arms. Two items were dropped because of their outlying item parameter values and content for a final count of 79 items.

In the development of the MAT-sf, 10 items were selected from clusters B–G described previously using the following criteria: (a) selected items exhibited a reasonable spread of item difficulty, (b) there was at least one item from clusters B–G, and (c) items had a relatively high discrimination parameter. We omitted items from cluster A for consideration because these items involved walking with a cane and were developed for use in a computerized adapted version of the measure, a long range goal of our research on this topic.

Snapshots of the 10 items in the MAT-sf can be found in Figure 1; there was at least one item chosen from each of the clusters described previously, and as will be described later, the items vary in item difficulty. Respondents are allowed to watch each video clip as many times as they like to completely understand the demands of each task by clicking a run and stop button; however, they are required to watch each video clip at least once before clicking their response to each of the 10 items (see Appendix). The MAT-sf test takes \(< 5\) minutes to complete. Note that Items 1 and 2 ask about walking and jogging with a response scale from none to \( 60 \) minutes spaced at \( 5\)-minute intervals. Items 3 and 4 involve walking up an inclined ramp with and without a handrail with possible responses being 0, 1, 2, 3, or 4 times. Finally, Items 5–10 involve walking while stepping over hurdles, walking uphill over uneven terrain, and climbing stairs that vary in demand with Items 9 and 10, including a bag or bags in hand while climbing. The response options to these latter six items are either no = 0 or yes = 1.

After examining the distribution of responses to the scales for the first four items and to accommodate the item calibration procedure, we aggregated responses and recoded them as follows: for Items 1 and 2, none = 0, 5–15 minutes = 1, 20–30 minutes = 2, and >30 minutes = 3. For Items 3 and 4, none = 0, 1 = 1, 2 = 2, and either 3 or 4 = 3. Based on an application of the Bayes theorem, an IRT-based score is automatically calculated by the program MULTILOG for each response pattern, treating the item parameters obtained during the calibration state as fixed. The scores have been standardized for interpretation with a \( M = 50 \) and a \( SD = 10 \).

**Reliability and Validity**

In order to evaluate test–retest reliability of the MAT-sf, 30 older adults were tested on two different occasions separated by an interval of 2 weeks. In addition, several procedures were employed to provide preliminary evidence for the validity of this measure. First, we calculated correlation coefficients between the sum of the total 79 items and scores on the 10-item MAT-sf. Second, we examined convergent validity of the MAT-sf by computing correlations between the MAT-sf and another index of self-reported disability, the PAT-D. Third, we conducted a \( t \) test on MAT-sf scores for older adults who completed versus those who failed the 400-m walk test. Finally, we gathered additional information related to the construct validity of the MAT-sf by examining the relationship of this measure to two performance-based measures of physical function, the SPPB, and gait speed during the 400-m walk test.
Figure 1. Snapshots of the 10 items and their response scales for the MAT-sf (instructions for downloading the computer program can be found in the research note at the end of the manuscript).
Measures

Demographic variables and chronic disease health status.—The age, sex, race, and presence of chronic disease for each participant were assessed via self-report. The presence of any chronic disease was coded as 1 and the absence coded as 0.

The short physical performance battery.—The SPPB involves timed measures of lower extremity performance: balance, chair stands, and 4-m self-paced walking speed (6). Performance in each of these three areas is assigned a categorical score ranging from 0 to 4, with 4 indicating the highest level of performance and 0 an inability to complete the test. A summary score ranging from 0 (worst performers) to 12 (best performers) is calculated by adding walking speed, chair stands, and balance scores.

Four hundred-meter walk test.—The 400-m walk test is a modified version of a self-paced walk test that was developed by Newman and her colleagues (5). In the modified version, participants are instructed to walk as quickly as they can for 400 m. Individuals walked 10 laps in a corridor between two cones spaced 20-m apart. The maximum time allowed for the test is 15 minutes. Time to complete the 400-m walk was recorded in minutes and seconds.

Pepper assessment tool for disability.—The PAT-D consists of 19 items that yield three subscales and a total score. The three subscales include ADL disability, mobility disability, and IADL disability. All factors have acceptable internal consistency reliability (>.70) and test–retest (.70) reliability coefficients. Fast walkers self-report better function on the PAT-D scales than slow walkers—effect sizes ranging from moderate to large (0.41–0.95); individuals with cardiovascular disease (CVD) have poorer scores on all scales than those free of CVD. In an 18-month randomized controlled trial, individuals who received a lifestyle intervention for weight loss had greater improvements in their mobility disability scores than those in a control condition (8).

Results

As shown in Table 1, participants had a mean age of 81.9 years. As expected for this age group, women represented just more than two thirds of the sample. The most common comorbidity was high blood pressure followed by arthritis and heart disease. Table 2 provides the descriptive statistics for the self-report and performance-based measures of disability used in the validation analyses along with data for the MAT-sf. In examining the means and standard deviations for the SPPB and 400-m walk test, it is clear that the sample had evidence of physical disability, yet exhibited considerable variability in physical function. This is most evident in the 400-m walk test where the average gait speed was 1.04 m/s, yet performance ranged from 0.45 to 1.76 m/s. Also, the range in the SPPB was from 2 to 12.

Content Validity

In the development of the MAT-sf, content validity was of central concern. Recall that the goal was to have a set of items that sampled the following six clusters (B–G): walking on a flat surface, walking up inclined ramps, walking while stepping over hurdles, walking outdoors uphill on uneven terrain, climbing stairs with and without handrails, and climbing stairs while carrying bags in one or both arms. In addition, we wanted to have a set of items for the MAT-sf that captured a broad range of abilities and items that provided valuable information to the measure.

Figure 2 presents the item characteristic curves (ICC; denoted by solid lines) and information curves (denoted by dotted lines) for each of the 10 items from the MULTILOG calibration procedure. Note that Items 5–10 that have a dichotomous no/yes response have only a single ICC; for Items 1–4 that have four categories of ability, the number of response curves is equal to the number of responses or 4. The x-axis for each graph describes the ability level that is being assessed by each item, whereas the y-axis describes the probability of a positive response to the item or category.
Figure 2. The item response and information curves for each of the 10 items in the MAT-sf. For Items 5–10, the response scale was dichotomous: no or yes. Thus, there is only a single curve needed to describe the ability level for that item. For Items 1 and 2, there are four categorical responses; for Items 1 and 2, none = 0 (yellow line), 5–15 minutes = 1 (green line), 20–30 minutes = 2 (blue line), and >30 minutes = 3 (red line); and for Items 3 and 4, none = 0 (yellow line), 1 = 1 (green line), 2 = 2 (blue line), and either 3 or 4 = 3 (red line).
in question. Hence, for Items 5–10 that have a dichotomous response, Item 7 is the easiest item and Item 10 is the most difficult. Note that Items 1–4, which have multiple response options, capture a range of abilities. A steeper ICC generally suggests higher discriminating power of the item or item category at the location where the curve has its steepest slope. On the other hand, the information curve depicted by the dotted line in each graph indicates the amount of information contained in each item. Higher information suggests more accurate estimates of ability for a particular item or category. In selecting the final 10 items for the MAT-sf, a goal was to have at least one item from clusters B through G described previously. In addition, we chose items that reflected a wide range of ability.

The 10 items cover the range of functional ability quite well. For example, a sequential examination of Items 1, 7, 8, 9, and 2 reveals a graduated increase in the complexity and difficulty that is inherent in different forms of mobility; these items also cover a broad range of the ability continuum. Note that both Items 3 and 4 tap a focal point in the middle section of the ability distribution. The trade-off is that, despite this narrow range, Items 3 and 4 have relatively high information content, which tends to stretch the information scale for all items. In addition, although it is beyond the scope of this article to discuss in detail the sensitivity of the animated video clips to contextual features of task demands, we would like to make two points. First, when respondents viewed a video clip of an animated figure walk at a slow (0.6 m/s) versus fast (1.3 m/s) pace, the percentage of participants who reported that they could do so for >30 minutes was 41% and 33%, respectively. When asked about climbing three stairs, carrying a light bag, either using or not using a handrail, the percentage of participants responding yes was 90% and 54%, respectively. Thus, it seems clear that providing more detailed information about the specific demands of task performance is critical to the assessment of mobility.

Reliability and Validity

Having identified the items for the MAT-sf, we then examined the reliability and validity of the measure. Because we had complete data on all 79 items, we began by calculating a composite score for each participant using all items and then correlated it with scores from the 10-item MAT-sf. As desired, the two were very highly related with one another; \( r = .96, p < .001 \). In addition, we conducted a 2-week test–retest reliability coefficient for the MAT-sf in a subsample of 30 participants and found that the measure was very stable over this time period, \( ICC = 0.93, p < .0001 \).

Several analyses were conducted to evaluate the validity of the MAT-sf. As support for convergent validity, we computed bivariate correlations between the MAT-sf and a validated self-report measure of disability, the PAT-D. Our hypothesis was that of the three PAT-D subscales—ADL, mobility, and IADL—the strongest relationship would exist between the PAT-D mobility subscale and the MAT-sf, whereas the weakest relationship would be found for the IADL subscale. This is exactly what occurred in that the correlations of the MAT-sf with the mobility, ADL, and IADL subscales of the PAT-D were \(-.60, -.50, \) and \(-.44, \) respectively; all correlations were significant \((p < .001)\). Using a \( t \) test for unequal variances, we also conducted a known groups test of construct validity finding that older adults who were able to complete the 400-m walk test have higher MAT-sf scores than those who failed the test; 56.09 (+/-70) versus 42.90 (+/-26), \( t(60.12) = 9.14, p < .0001 \).

Finally, as a further test of the construct validity of the MAT-sf, we conducted two separate stepwise regression analyses, one for the SPPB and a second for the 400-m walk test. In these analyses, the PAT-D mobility score was entered first followed by the MAT-sf score. In both analyses, the entry of the MAT-sf contributed over and above the PAT-D mobility subscale to the explanation of performance-based function; for the SPPB, the incremental change in \( R^2 \) was 9.8%, and for the 400-m walk test, it was 16.7%. The bivariate correlations of the MAT-sf with the SPPB and 400-m walk test were \(.59 (p < .001)\) and \(.58 (p < .001)\), respectively. It is also of interest to point out the standardized \( \beta \) weight for the MAT-sf was substantially larger than the PAT-D mobility subscale in both analyses (Table 3).

### Table 3. Final Regression Models of PAT-D Mobility and MAT-sf on the SPPB and 400-m Walk Gait Speed

<table>
<thead>
<tr>
<th>Measure</th>
<th>Standardized ( \beta )</th>
<th>( t ) Value</th>
<th>( p ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPPB, ( R^2 = 40% )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAT-D mobility</td>
<td>(-.31)</td>
<td>(-4.78)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>MAT-sf</td>
<td>0.40</td>
<td>6.11</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>400-m walk gait speed, ( R^2 = 37% )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAT-D mobility</td>
<td>(-.20)</td>
<td>(-2.91)</td>
<td>.004</td>
</tr>
<tr>
<td>MAT-sf</td>
<td>0.48</td>
<td>6.91</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Notes: MAT-sf = Short form of Mobility Assessment Tool; PAT-D = Pepper Assessment Tool for Disability; SPPB = Short Physical Performance Battery.

**Discussion**

The MAT-sf that uses animated video presentation of content had excellent content validity. Using state-of-the-art psychometric methods (9,10) on a pool of 79 items, we were able to identify a subset of 10 items that represented a reasonable universe of content for mobility, provided good discrimination across a range of abilities, and had other favorable psychometric characteristics. That is, the MAT-sf had acceptable test-retest reliability and shared common variance with an existing measure of self-reported disability, the PAT-D. Of interest is the fact that the MAT-sf correlated strongest with the PAT-D mobility subscale and weakest with the PAT-D IADL subscale, providing evidence that the measure has both convergent and discriminant validity. Finally, scores on the MAT-sf were lower for those who failed versus those who were successful in completing the 400-m walk test—41% and 33% respectively; all correlations were significant \((p < .001)\).
walk test, evidence of known groups construct validity (12). Failure to complete the 400-m walk test is an excellent marker of mobility disability (13), and the loss of mobility is predictive of multiple adverse events, including morbidity, worsening of mobility disability (e.g., ADL disability), institutionalization, and mortality (3,14–16). More important, after entry of the PAT-D mobility subscale in a regression analysis, the MAT-sf explained additional variance in both SPPB scores and the 400-m walk test. The bivariate correlations of the MAT-sf with both SPPB and the 400-m walk test represent large effect sizes.

The MAT-sf provides a brief measure of mobility that can be completed in less than 5 minutes. Whereas it has obvious utility in large epidemiological studies and randomized controlled trials that have mobility as an outcome, we also believe that this instrument may be useful for translational scientists. For example, we are in the process of determining whether the MAT-sf can be used in a variety of clinical contexts to help geriatricians assess and advise older adults about the status of their mobility. In essence, we view mobility to be an important vital sign in treating older adults, and the MAT-sf might provide a pragmatic solution to mobility assessment in the context of primary care.

A current limitation of the MAT-sf is that data do not exist on sensitivity to change as a function of physical activity or drug-related treatments that might be employed to intervene on mobility. At this point, we also do not know how scores on the MAT-sf will predict adverse events, such as morbidity, ADL disability, institutionalization, and mortality. In addition, large-scale administration of the MAT-sf is needed to provide normative data on the measure. Projects are currently in progress or in the planning stages to address these deficiencies. Finally, there are those who would argue that a brief performance-based index of mobility would be more informative. Our position is that although performance-based measures of function provide valuable information on what a person can do, older adults’ perceptions of their abilities are important determinants of what they will do in their daily lives. In other words, performance and self-reported measures of mobility not only share common variance but also offer unique information as researchers seek to better understand the process of physical disablement.

To summarize, the goal of this research was to develop a brief psychometrically sound measure of mobility that standardized representation of task performance along dimensions of speed, the slope of inclines, and other environmental challenges. We found that this new measure, the MAT-sf, improved upon the prediction of performance-based measures of function when contrasted with a more traditional assessment tool, the PAT-D. The use of animation in the video clips helped to avoid biases that would have been introduced by using men and women with different cultural and morphologic characteristics. We want to underscore the fact that we had no intent to test the incremental value of video technology over and above either an oral or a pen and paper format of identical item content. Indeed, other investigators may want to ask the question of whether complex oral or written formats may work as well as the stimuli used in the MAT-sf; however, this is a formidable challenge particularly given test administration time constraints in large epidemiological studies and when working with older adults who have compromised attention and/or comprehension. At this point, we are simply stating that the MAT-sf is psychometrically sound and that the content enables researchers to capture the complexity inherent in mobility in a manner that is innovative and of value to researchers in the fields of gerontology and geriatric medicine.

**Funding**

Support for this study was provided by (a) National Institutes for Aging P30 AG021332, (b) National Heart, Lung, and Blood Institute grant HL076441-01A1, and (c) General Clinical Research Center grant 5M01RR007122-18.

**References**

APPENDIX:
Information on the short form of the mobility assessment tool (MAT-sf):
• Although the manuscript includes a figure to illustrate the items, to see a demo of the measure, go to the following website: http://mat-sf.wfuhs.arane.us/.
• A Windows version of the MAT-sf application can be directly downloaded from http://mat-sf.wfuhs.arane.us/MCAT10S-TestApplication/apps/MAT.exe.
• MAT-sf requires up-to-date versions of Java and QuickTime, which may be downloaded from the following two URLs: http://www.java.com/en/download/; http://www.apple.com/quicktime/download/.
• The MAT.exe executable file may be placed anywhere on the hard drive, including the desktop, if desired.
• Depending on the speed of the computer, it may take up to a minute between attempting to run the MAT.exe file and the program appearing.
• If you have problems or questions about the program, E-mail our computer programmer, Ryan Barnard at rybarnar@wfubmc.edu.