Published in final edited form as:

Qual Life Res. 2013 October ; 22(8): 1907–1915. doi:10.1007/s11136-012-0346-9.

Psychometric properties of a computerized adaptive test for assessing mobility in older adults using novel video-animation technology

Edward H. Ip,

Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd., WC23, Winston-Salem, NC 27157, USA

W. Jack Rejeski,

Department of Health and Exercise Science, Wake Forest University, Winston-Salem, NC, USA

Anthony P. Marsh,

Department of Health and Exercise Science, Wake Forest University, Winston-Salem, NC, USA

Ryan Barnard, and

Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd., WC23, Winston-Salem, NC 27157, USA

Shyh-Huei Chen

Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd., WC23, Winston-Salem, NC 27157, USA

Edward H. Ip: eip@wfubmc.edu

Abstract

Purpose—This paper reports on the psychometric properties of a computerized adaptive test (CAT) version of the Mobility Assessment Tool (MAT) for older adults (MAT-CAT).

Methods—An item pool of 78 video-animation-based items for mobility was developed, and response data were collected from a sample of 234 participants aged 65–90 years. The video-animation-based instrument was designed to minimize ambiguity in the presentation of task demands. In addition to evaluating traditional psychometric properties including dimensionality, differential item functioning (DIF), and local dependence, we extensively tested the performance of several MAT-CAT measures and compared their performances with a fixed format.

Results—Operationally, the MAT-CAT was sufficiently unidimensional and had acceptable levels of local independence. One DIF item was removed. Most importantly, the CAT measures showed that even starting with a single fixed item at the mean ability, the adaptive version delivered better performance than the fixed format in terms of several criteria including the standard error of estimate.

Conclusion—The MAT-CAT demonstrated satisfactory psychometric properties and superior performance to a fixed format. The video-animation-based adaptive instrument can be used for assessing mobility with specificity and precision.

Keywords

Mobility Assessment Tool; Item response theory; Health-related quality of life; Mobility disability; Animation

Introduction

In the fields of medicine and gerontology, health-related quality of life (HRQOL) is often used to describe a broad range of patient-reported outcomes, of which physical functioning is a critical component [1]. Indeed, the loss of functional independence, especially difficulty with mobility, is a harbinger of further health complications. Performance-based assessment of mobility, while possible in some circumstances, is not always feasible. Furthermore, performance-based tests such as the Short Physical Performance Battery (SPPB) [2] and locomotion tasks such as the 400-m walk [3] convey different information than people's perceptions of their physical capacities. Although several measures of self-reported physical functioning do exist [4–6], the development of advanced computer-based methods for measuring patient-reported outcomes (PROs) for mobility and physical functioning has been undertaken in the PRO Measurement Information System (PROMIS) network [7].

Traditional self-report measures that assess mobility and other facets of physical functioning have known limitations. A common challenge is that measures require participants to make complex judgments about the implicit meaning inherent in written descriptions of various tasks. Contextual factors such as the environment in which the behavior takes place are ignored, complicating an individual's judgment about task difficulty. For example, when asking about limitations in walking up a flight of stairs, one needs to consider factors such as how many steps, how fast, and can a handrail be used? The lack of a common interpretation of tasks could lead to substantial response biases.

Marsh et al. [8] provided evidence that varying the contextual features and demands of a simple task such as stair climbing (e.g., with and without a handrail) and walking (e.g., at 0.6, 1.0, and 1.3 m/s) have a significant impact on older adults' perceptions of their abilities. Using a video-animation-based presentation from the Mobility Assessment Tool (MAT), the authors found that when asked, "Can you walk up 3 stairs, using a handrail, at the pace shown?", the percentage of participants who reported an inability to do the task were only 1 %. When participants were shown a video animation without a handrail, the percentage reporting an inability increased to 22 %, underscoring the importance of this contextual feature.

A short form of the MAT consisting of 10 items, the MAT-sf, has been developed and validated; the results have been reported elsewhere [9]. The purpose of the current paper is to extend this work by describing the psychometric properties and advantages of the computerized adaptive test (CAT) version of MAT, or MAT-CAT, which employs an item bank of 81 animated video clips. As noted above, the most important advantage of using

video-based assessment is the significant improvement in the level of specificity provided to the respondent. In addition to specificity, there is also evidence, suggesting that video-based assessments have a higher level of acceptability, especially among respondents with low levels of literacy [10]. Although the MAT-sf is a valid and reliable approach to assessing mobility in older adults, our hypothesis is that the MAT-CAT will have similar psychometric properties yet exhibit superior performance (to be defined later) to the MAT-sf.

Methods

Participants and recruitment

A total of $N = 234$ men and women were recruited from independent living residences that were part of four older adult communities managed by Senior Living Communities (senior-living-communities.com) in North and South Carolina. Participants were recruited by local social/wellness directors and through lectures on aging given by senior investigators on the project. The methods for this investigation were approved by an institutional board for human subject research, and all participants completed an informed consent form prior to participation in the data collection. The inclusion criteria included: aged 65–90 years, ability to walk without help either with or without the use of an assistive device, a Mini-Mental State Exam (MMSE) ≥ 23 , and willingness to sign informed consent and Health Insurance Portability and Accountability Act (HIPAA) authorization forms. Participants were excluded from the study if they had one or more of the following conditions: undergoing active treatment for a psychiatric illness, severe symptomatic heart disease, resting blood pressure $>160/100$ mmHg, or severe systemic diseases. Table 1 shows the demographic characteristics and pertinent clinical variables for the sample. In order to evaluate the test–retest reliability of the MAT, $N = 30$ older adults were tested on two different occasions separated by an interval of 2 weeks. Data were collected through self-contained programs that were downloadable to individual machines.

Measures

MAT item pool and MAT-sf—The development of the MAT scale was based on an item bank of 81 animated video clips for mobility tasks, which were categorized into seven functional clusters (# of items): Cluster A (10)—ambulating at various speeds 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 at different speeds, Cluster D (2)—walking while stepping over hurdles, Cluster E (6)—walking outdoor uphill on uneven terrain with different inclines and at different speeds, 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 of different runs. Three items were dropped from the item pool because of their poor fit to the data and unsatisfactory psychometric properties (to be described later), which left a final bank of 78 items.

Based on the sample of $N = 234$ older adults, the MAT items were calibrated using item response theory (IRT) [8]. When creating the short-form version of the MAT (MAT-sf) [9], the intraclass correlation (ICC) and goodness-of-fit statistics of the items were carefully

examined, and 10 items were selected from Clusters B–G described previously using the following criteria: (1) selected items represent a broad range of mobility tasks, (2) items exhibited a reasonable spread of item difficulty, (3) items exhibited a high level of goodness of fit to the data and reliability, and (4) most items had a relatively high discrimination parameter. Cluster A items, which involved walking with a cane, were not included in MAT-sf, but these items were used in the MAT-CAT.

Figure 1 shows screenshots of 10 representative items in the MAT. At least one item was chosen from each cluster previously described. Item 6 asks about walking with a cane. Item 11 asks about walking and jogging with a response scale from none to 60 min spaced at 5-min intervals. Items 16 and 19 involve walking up an inclined ramp with and without using a handrail with possible responses being 0, 1, 2, 3, or 4 times. The last 6 items involve walking while stepping over low hurdles, walking uphill over uneven terrain, climbing stairs that vary in demand, carrying bags in one hand while climbing stairs, and carrying bags in both hands while climbing stairs, respectively.

The item pool contains both polytomous ($n = 20$) and dichotomous ($n = 61$) items. In the previous report [8], the calibration used both polytomous and dichotomous items; however, in the current study, we treated all items using binary responses.

MAT-CAT measures—MAT-CAT measures were derived from an analysis of participants' responses to the entire item pool. The underlying engine for MAT-CAT was based on an iterative algorithm that administered a set of items to individuals that was tailored to his or her level of mobility ability [11]. Based on a current estimate of ability, the adaptive algorithm searched the item pool for the optimal item and then administered it to the individual. The individual's ability estimate was then updated by incorporating the information received from the response to the item. In this study, we used the maximum Fisher information (MFI), a commonly used criterion, for selecting optimal items. The item with MFI is one that would lead to the largest reduction in the confidence interval of the ability estimate.

Two important issues concerning the practical implementation of MAT-CAT are as follows: (1) determination of initial ability estimates and (2) the stopping criterion. To evaluate different schemes for implementing (1) and (2), we used four different initial conditions and four stopping criteria. The conditions for assigning the initial estimate for MAT-CAT were specified as follows:

1. CAT0: A single item with MFI at ability = 0,
2. CAT1: A single item with MFI at ability = -0.5 (0.5), assuming that prior knowledge is available confirming that ability > 0 (> 0),
3. CAT2: Two items with MFI by specifying ability = ± 0.5 ,
4. CAT3: Including all the three items in CAT0 and CAT2.

The four stopping criteria for MAT-CAT were fixed test lengths of 3, 5, 7, and 10 items. We did not implement a CAT procedure of variable length. All MAT-CAT measures were derived using a customized program written in the R language [12, 13] by the research team.

Short physical performance battery (SPPB)—The SPPB is a commonly used measure of physical functioning in population aging studies and is composed of three tasks: a hierarchical balance task, a short walk at the usual speed, and five repetitive chair stands [2]. A summary score ranges from 0 (worst) to 12 (best).

Pepper assessment tool for disability (PAT-D)—The PAT-D is a validated instrument [6] of self-reported disability, and the test consists of 19 items that yield three subscales (ADL disability, mobility disability, and IADL disability) and a total score.

Four hundred-meter walk test—Participants were instructed to walk as quickly as they can for 400 m in a corridor between two cones spaced 20 m apart for 10 laps. The maximum time allowed for the test was 15 min. Time to complete the 400-m walk was recorded in minutes and seconds.

Statistical analysis

The calibration procedure for the items has been reported in earlier articles [8]. Psychometric analysis followed the procedure described in Reeve et al. [7] and consists of the following components: (1) evaluation according to traditional descriptive statistics including item-level response distribution, as well as ceiling and floor effects, internal consistency, reliability, and validity; (2) evaluation of assumptions of IRT, including dimensionality analysis and local independence; (3) differential item functioning (DIF) evaluation; and (4) evaluation of the performance of the MAT-CAT measures.

Descriptive statistics—We examined response frequency and scale statistics, including the distribution of item response, internal consistency reliability coefficients, test–retest reliability, and correlations of the measure with the SPPB, the 400-m walk, and the PAT-D.

Assumptions of IRT—To evaluate the assumption of dimensionality, we used confirmatory factor analysis (CFA) as well as bifactor analysis [14]. Bifactor analysis posits that for each item, there first exists a common underlying dimension, and that there also exists a domain-specific dimension for the individual item belonging to the domain. The bifactor analysis was used to affirm the presence of the clusters A–F. We reported the $-2 \times \log\text{likelihood}$, the Akaike's Information Criterion (AIC), and the Bayesian Information Criterion (BIC) of different fitted models. We also reported model-fit indices: the Tucker–Lewis Index (TLI), the Comparative Fit Index (CFI), and root-mean-squared error of approximation (RMSEA). The following criteria were used to determine the goodness of fit of the unidimensional model: TLI = 0.9, CFI = 0.95, RMSEA = 0.6. The goodness-of-fit (GFI) index was not included because of recent work regarding its lack of sensitivity [15]. For the evaluation of the local independence assumption, we used Yen's Q_3 statistic [16], which is the pairwise correlation between residuals of item responses after partialling out the underlying factor. The evaluation of local independence involved testing the significance of the residual correlation on a large number of item pairs (in our case 3,003 pairs). To control for the multiplicity in comparison, we first applied an r -to- z trans-form to the Q_3 statistic, and then, we used the Hochberg's procedure [17], which uses a sliding scale for testing p

values and is generally considered a more powerful procedure than the widely used Bonferroni's adjustment. The nominal level of significance was set at $\alpha = 0.05$.

Differential item functioning—DIF was evaluated using the Wald's test for determining whether the parameters show signs of drifting across different group memberships. The Wald's test assesses statistical significance in the chi square values in both the discrimination and the difficulty parameters in the IRT model. In other words, both uniform DIF and non-uniform DIF were tested [18]. Because of the large number of items involved, we used the Hochberg's procedure to evaluate significance with a nominal level of significance set at $\alpha = 0.05$. For this report, we only included the variable gender in the DIF analysis. Race was not included because the sample was predominantly White and contained only 8 % Black participants (Table 1).

Item calibration and scoring were conducted using IRTPRO 2.1 (Scientific Software International, Inc.) [19]. Confirmatory factor analysis, bifactor analysis, and DIF analysis were also conducted using IRTPRO. In addition, we used Mplus [20] to supplement our toolbox and to derive commonly used model-fit statistics in CFA-TLI, CFI, and RMSEA.

Evaluation of MAT-CAT measures—We used several performance outcomes to evaluate the MAT-CAT measures. We studied the relationship between the number of items of the test and the ceiling and floor effects. Furthermore, we studied the correlation between the score derived from CAT and the score derived from the full version as a function of test length across CAT0 to CAT3. Other metrics used in evaluating the performance of the different CAT measures included the root-mean-squared error (RMSE) and the standard error (SE) of ability estimate, which, respectively, measures bias and accuracy of the ability estimate. Throughout the analysis, we used a fixed-format version of MAT—the MAT-ff, which contains the same 10 items from the short-form MAT-sf. The only difference is that in MAT-ff, all 10 items were dichotomously scored.

Results

Descriptive statistics

The descriptive statistics for measures of disability used in the validation analyses are presented in Table 2. Of the 78 dichotomized MAT items, there were 14 participants (5.9 %) who responded positively to all binary items, but there were none who responded negatively to all items. In contrast, the MAT-ff had 8.5 % participants who hit the ceiling and 5.1 % who hit the floor. The ceiling and floor effects for the MAT-CAT measures are reported in the section regarding MAT-CAT assessment. The mean correlation between an item and the total score with the item deleted was 0.622 (SD = 0.148), and the Cronbach's alpha was 0.981, suggesting a high degree of internal consistency among the items. The value of intraclass correlation (ICC) for test-retest reliability was 0.97. Correlations between the MAT score (higher better) and the SPPB, 400-m walk time, and PAT-D (lower better) were 0.60, −0.40, and −0.67, respectively, and the correlations for the MAT-ff were 0.58, −0.38, and −0.56, respectively.

Assumptions of IRT

Table 3 summarizes the goodness-of-fit indices for the CFA (one-dimensional and two-dimensional IRT) and bi-factor analysis of MAT data. It shows that the bifactor (7 factors: 1 general and 6 representing item clusters A–F) model had the lowest AIC, BIC, and loglikelihood ratio. However, the goodness-of-fit statistics between the bifactor model and the other two models cannot be directly compared because they are not nested models. The model-fit statistics for the one-dimensional IRT model, based on mean and variance-adjusted weighted least squares (WLSMV) estimation, are as follows: TLI = 0.985, CFI = 0.975, and RMSEA = 0.126. All indices show acceptable fit. For the two-dimensional IRT and bifactor analysis, some parameter values were found to be rather extreme. There were approximately 3.3 % of all item pairs that significantly deviated from the local independence assumption after adjustment for multiple comparisons using the Hochberg's procedure. The unadjusted percentage of local, dependent item pairs were much higher, at approximately 21 %. Thus, there is some evidence that points to violation of the local independence assumption. A closer examination of the item pairs, however, revealed that the locally dependent items tended to cluster around the same task (e.g., walking slower and walking faster).

Differential item functioning

Eight items showed significant DIF for both the discrimination and the difficulty parameters at the unadjusted nominal level of $\alpha = 0.05$; however, after applying the Hochberg's procedure [17, 21] for adjusting for multiple comparisons, only one item remained significant. A closer examination of the item (Item 28: walk-up 3 stairs using handrail at moderate pace) indicated that its parameter values were very high. The item was excluded from the item pool and subsequent CAT analysis.

Evaluation of MAT-CAT measures

For CAT0, the item with MFI at an ability value of 0 was Item 38 (walk-up 6 stairs without using handrail at a moderate pace). Figure 2 shows the trajectories of ability estimates for the CAT0 implementation of a random sample of 23 participants. It can be seen that the estimate begins to stabilize after administering approximately 10 items, after which the gain in accuracy is marginal.

Table 4 shows the ceiling and floor effects for MAT-CAT implementations of varying lengths (3, 5, 7, and 10 items). This table can be compared with the full-version MAT and the MAT-ff—all four 10-item CAT implementations have lower ceiling and floor effects than the 10-item MAT-ff.

Using MAT-CAT measures of up to 10 items, we show the Pearson's correlation coefficients between the CAT0, CAT1, CAT2, CAT3, MAT-ff, and the full-item bank in Fig. 3. Not surprisingly, the Pearson's coefficients for CAT0, CAT1, CAT2, CAT3, and the full-item bank all increase with the number of items administered in MAT-CAT. The coefficients for all MAT-CAT measures are higher than the non-adaptive MAT-ff after approximately 5 items. Figure 4 shows the RMSEs for the various MAT-CAT measures. It

further supports the result that after 5 items, the performance of MAT-CAT is better than that of the MAT-ff.

Figure 5 shows the standard errors (SEs) of ability estimates for different lengths of the MAT-CAT measures. The SEs of 5-, 7-, and 10-item MAT-CAT are compared with the MAT-ff and the full version of MAT, and all curves exhibit the U-shaped pattern typically seen in plots of SE against ability. Because width of confidence interval is proportional to SE, the pattern in Fig. 5 would remain unchanged if width of confidence interval of the ability estimate is used instead of SE. Compared with other MAT measures, including the MAT-ff, the 10-item CAT0 demonstrates good precision, especially at the upper extremity of the ability spectrum. For example, at 1 standard deviation above the mean, reliability of the estimate for CAT0 is approximately 0.86 ($SE = 0.2$, from Fig. 5), whereas reliability for MAT-ff is approximately 0.69 ($SE = 0.4$).

Discussion

This article focuses on the psychometric properties of a CAT version of the MAT or MAT-CAT, a video-animation-based tool for assessing mobility in older adults. The evaluation of the MAT-CAT led to several important findings: (1) The performance of the CAT version, regardless of the initial starting condition, is better than the non-adaptive, 10-item, fixed-format MAT-ff; the gain in performance starts to manifest at approximately 6 adaptive items. (2) The performance of a 10-item adaptive test is rather independent of how the initial estimate is selected—starting at 1 item at the average ability is almost as good as starting with multiple pre-specified items. And (3), the gain in performance becomes marginal after 10 items; therefore, we would not recommend administering more than 10 items in MAT-CAT.

Interestingly, we found that MAT-CAT is significantly more efficient than MAT-ff. At 6 items, the MAT-CAT already demonstrated correlations of equal magnitude with the full version of the MAT (correlation = 0.96), as well as smaller RMSE than the MAT-ff (0.35 for CAT0 vs. 0.40 for MAT-ff). The contrast between MAT-CAT and MAT-ff is especially marked at the higher end of the mobility–ability continuum. Figure 5 attests to this observation. Judging by the SE of the estimate, MAT-ff has its limitations, and the gap from the full version of the MAT is especially salient for ability values in the range between 0.5 and 1.0 standard deviations above the mean.

Another finding is related to model fitting. We have fitted multiple models to the data, and the dimensionality analysis suggested that a unidimensional model offers acceptable fit to the data using several goodness-of-fit indices. For other goodness-of-fit indices including AIC and BIC, more complex models—such as the 7-factor bi-factor model—appeared to be superior. However, a closer examination of the complex models revealed that (1) the values of some item parameters within these complex models were extreme and (2) the additional dimensions were all related to specific tasks and do not represent substantive and meaningful dimensions. We thus concluded that the data could be sufficiently represented by a parsimonious unidimensional model.

There are several limitations to this study. First, we have not used actual live CAT response data. The MAT-CAT data have been simulated using the full-response data set. While this makes comparison between the MAT-CAT measures and the full version convenient, it is not clear that participants would respond the same way in a live CAT implementation. Further work will be needed to collect live CAT data from participants. Second, we did not study the use of individualized stopping criteria that are based on the precision of the estimate for each individual—for example, the test stops when the SE of estimate of an individual drops below a pre-determined threshold. The investigation reported in this paper, however, suggests that the gain after 10 items will be minimal. Therefore, for practical purposes, the use of individualized stopping criterion may not lead to meaningful gains in precision. Furthermore, our experience with CAT is that sometimes the individualized method may not achieve the required precision even after a long test. Another limitation of the study is the lack of data for evaluating sensitivity and predictive validity of the instrument. Whether or not scores on the MAT-CAT will predict important events such as ADL disability and hospitalization, and mortality is not known. We suspect that MAT-CAT could be more sensitive to subtle changes in mobility because of its adaptive nature, but longitudinal data will be needed to test such hypothesis. It is interesting to note, however, that the absolute correlation of the MAT-CAT with the 400-m walk test was slightly higher ($r = -0.67$) than for the MAT-ff ($r = -0.56$). The MAT-CAT provides the ability to efficiently draw from a large item pool of diverse physical tasks that are wide ranging in difficulty and holds the promise of allowing the evaluation of groups of older adults who exhibit a broad range of functional abilities. Finally, the limitation of the sample could affect the further generalization of our findings. For example, the sample is predominantly White and is relatively healthy. Currently, we are working on a larger sample of older adults to provide further validation of the instrument.

In summary, we demonstrate in this report that the MAT-CAT has sound psychometric properties and could be used by HRQOL researchers to study mobility in older adults. Future research should focus on an actual CAT implementation, study more diverse groups of older adults, and examine the ability of MAT-CAT to predict health outcomes such as nursing home admissions, falls, and rate of disease progression.

Acknowledgments

The research has been supported by the following grants from the National Institute of Aging: P30 AG21332.

Abbreviations

DIF	Differential item functioning
MAT	Mobility Assessment Tool
HRQOL	Health-related quality of life
PRO	Patient-reported outcome
HIPAA	Health Insurance Portability and Accountability Act of 1996
CAT	Computerized adaptive test

References

1. Rejeski WJ, Mihalko S. Physical activity and quality of life in older adults. *The Journals of Gerontology*. 2001; 56A(Special Issue II):23–35. [PubMed: 11730235]
2. Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *The Journals of Gerontology*. 1994; 49:M85–M94.
3. Newman AB, Simonsick EM, Naydeck BL, et al. Association of long-distance corridor walk performance with mortality, cardiovascular disease, mobility limitation, and disability. *JAMA*. 2006; 295(17):2018–2026. [PubMed: 16670410]
4. Ware, JE.; Kosinski, M.; Keller, SK. SF-36 physical and mental health summary scales: A user's manual. Boston, MA: The Health Institute; 1994.
5. Cella DF, Tulsky DS, Gray G, et al. The functional assessment of cancer therapy scale: Development and validation of the general measure. *Journal of Clinical Oncology*. 1993; 11(3): 570–579. [PubMed: 8445433]
6. Rejeski W, Ip E, Marsh A, Miller M, Farmer D. Measuring disability in older adults: The ICF framework. *Geriatrics & Gerontology International*. 2008; 8(1):48–54. [PubMed: 18713189]
7. Reeve BB, Hays RD, Bjorner JB, et al. Psycho-metric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*. 2007; 45:S22–S31. [PubMed: 17443115]
8. Marsh AP, Ip EH, Barnard RT, Wong YL, Rejeski WJ. Using video animation to assess mobility in older adults. *Journals of Gerontology Series A, Biological Sciences and Medical Sciences*. 2011; 66(2):217–227.
9. Rejeski WJ, Ip EH, Marsh AP, Barnard RT. Development and validation of a video-animated tool for assessing mobility. *Journals of Gerontology Series A, Biological Sciences and Medical Sciences*. 2010; 65(6):664–671.
10. Curcio, CL.; Gómez, JF.; Lord, C.; Ip, EH.; Marsh, AP.; Rejeski, WJ.; Alvarado, BE. Validity and reliability of Spanish version of the Mobility Assessment Tool short version: Results from a Colombian study in elderly populations. Presented at the 2011 Canadian Association on Gerontology Pan-American Congress on Geriatrics and Gerontology; Ottawa, Canada. October 2011; 2011.
11. Wainer, H., et al., editors. Computerized adaptive testing: A primer. 2. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
12. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2011. <http://www.R-project.org/>
13. Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*. 2006; 17(5):1–25.
14. Gibbons RD, Hedeker DR. Full-information item bifactor analysis. *Psychometrika*. 1992; 57(3): 423–436.
15. Sharma S, Mukherjee S, Kumar A, Dillon WR. A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*. 2005; 58(1):935–943.
16. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*. 1984; 8(2):125–145.
17. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75(4):800–802.
18. Lord FM. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*. 1977; 1(1):95–100.
19. Cai, L.; Thissen, D.; du Toit, S. IRTPRO user's guide. Lincolnwood: Scientific Software International, IL; 2011.
20. Muthén, LK.; Muthén, BO. Mplus user's guide. 5. Los Angeles, CA: Muthén & Muthén; 2007.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*. 1995; 57(1):289–300.

Fig. 1.
A representative sample of the 78 items included in the mobility assessment tool item pool.
Each *cluster* contains items that are similar (e.g., walking at different speeds on a ramp)

Fig. 2. Ability estimates of a sample of 23 randomly selected participants from the computerized adaptive test implementation using a fixed starting item as a function of number of items administered

Fig. 3.
Correlation between various MAT-CAT measures and the full-item set. The correlation between MAT-ff and MAT is shown as a *horizontal dotted line*

Fig. 4. Root-mean-squared error (RMSE) of the various MAT-CAT measures. The RMSE of MAT-ff is shown as a *horizontal dotted line*

Fig. 5. Standard error (SE) estimates across the continuum of ability for full-length MAT (*solid line*), MAT-CAT measures of varying lengths, and fixed format (*FF*)

Table 1Characteristics of the sample of $N = 234$ of older adults

Characteristic	N (Mean $\pm$ SD or %)
Age (years)	238 (81.89 $\pm$ 5.25)
Race	
Black	19 (7.98 %)
White	219 (92.02 %)
Sex	
Male	68 (28.57 %)
Female	170 (71.43 %)
Education	
Elementary	26 (10.92 %)
High school	55 (23.11 %)
College	105 (44.12 %)
Post graduate	45 (18.91 %)
Other	7 (2.94 %)
Comorbidities	
Heart disease	77 (32.35 %)
High blood pressure	154 (64.71 %)
Diabetes	29 (12.24 %)
Arthritis	87 (43.07 %)

Table 2

Descriptive statistics for study measures

Measure	N (Mean $\pm$ SD)
400-m walk time (m/s)	187 (1.04 $\pm$ 0.27)
Short Physical Performance Battery	232 (8.61 $\pm$ 2.70)
Pepper assessment tool—disability	234 (0.37 $\pm$ 0.42)

Table 3

Goodness-of-fit statistics from confirmatory factor analysis (3 models: one-dimensional, two-dimensional, and 7-factor bifactor model) for assessing dimensionality of the MAT

	One-dimensional	Two-dimensional	Bifactor
–2loglikelihood*	10412.6	9819.6	9350.4
AIC*	10716.6	10275.6	9804.4
BIC*	11241.8	11063.4	10588.7

* Lower value represents better fit

Table 4

Floor and ceiling effects for various MAT-CAT measures

	3-Item (%)		5-Item (%)		7-Item (%)		10-Item (%)	
	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling
CAT0	21.4	28.2	6.0	15.8	2.6	14.5	1.3	6.8
CAT1	9.4	21.8	4.7	15.0	2.6	14.5	0.9	6.8
CAT2	13.7	28.6	6.0	15.8	3.0	14.5	1.3	6.8
CAT3	22.6	40.2	8.5	21.8	4.7	15.0	1.7	6.8