



Original Research

# Improving the Delivery of Function-Directed Care During Acute Hospitalizations: Methods to Develop and Validate the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT)



Andrea L. Cheville, MD, MSCE <sup>a</sup>, Chun Wang, PhD <sup>b</sup>,  
Kathleen J. Yost, PhD <sup>c</sup>, Jeanne A. Teresi, EdD, PhD <sup>d,e</sup>,  
Mildred Ramirez, PhD <sup>d</sup>, Katja Ocepek-Welikson, M Phil <sup>d</sup>,  
Pengsheng Ni, MD, MPH <sup>f</sup>,  
Elizabeth Marfeo, PhD, MPH, OTR/L <sup>g</sup>,  
Tamra Keeney, DPT, PhD <sup>h</sup>, Jeffrey R. Basford, MD, PhD <sup>a</sup>,  
David J. Weiss, PhD <sup>i</sup>

<sup>a</sup> Department of Physical Medicine and Rehabilitation, Mayo Clinic, Rochester, Minnesota

<sup>b</sup> College of Education, University of Washington, Seattle, Washington

<sup>c</sup> Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

<sup>d</sup> Research Division, Hebrew Home at Riverdale, Riverdale, New York

<sup>e</sup> Columbia University Stroud Center at New York State Psychiatric Institute, New York, New York

<sup>f</sup> School of Public Health, Boston University, Boston, Massachusetts

<sup>g</sup> Tufts University, Department of Occupational Therapy, Medford, Massachusetts

<sup>h</sup> Division of Palliative Care and Geriatric Medicine, Mongan Institute Center for Aging and Serious Illness, Massachusetts General Hospital, Boston, Massachusetts

<sup>i</sup> Department of Psychology, University of Minnesota, Minneapolis, Minnesota

*List of abbreviations:* AMC, Adaptive Measurement of Change; AM-PAC, Activity Measure of Post-Acute Care; CAT, computerized adaptive testing; DIF, differential item functioning; EHR, electronic health record; FAM, Functional Assessment for Acute Care Multidimensional; FAMCAT, Functional Assessment in Acute Care Multidimensional Computer Adaptive Test; HIPAA, Health Insurance Portability and Accountability Act of 1996; IRT, item response theory; MCAT, multidimensional computerized adaptive testing; MGRM, multidimensional graded response model; MIRT, multidimensional item response theory; PAC, postacute care; PH, physical function; PROM, patient-reported outcome measure; PROMIS, Patient-Reported Outcomes Measurement Information System; SF, short form.

This work was supported in part by the Mayo Clinic Kern Center for the Science of Healthcare Delivery; an Agency for Healthcare Research and Quality National Research Service Award T32 (grant #5T32 HS000011-33); and a Center on Health Services Training and Research fellowship funded by the Foundation for Physical Therapy Research.

Cite this article as: Arch Rehabil Res Clin Transl. 2021;3:100112

<https://doi.org/10.1016/j.arrct.2021.100112>

2590-1095/© 2021 The Authors. Published by Elsevier Inc. on behalf of American Congress of Rehabilitation Medicine. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**KEYWORDS**

Cognition;  
Activities of daily living;  
Rehabilitation

**Abstract Objective:** To (1) develop a patient-reported, multidomain functional assessment tool focused on medically ill patients in acute care settings; (2) characterize the measure's psychometric performance; and (3) establish clinically actionable score strata that link to easily implemented mobility preservation plans.

**Design:** This article describes the approach that our team pursued to develop and characterize this tool, the Functional Assessment in Acute Care Multidimensional Computer Adaptive Test (FAMCAT). Development involved a multistep process that included (1) expanding and refining existing item banks to optimize their salience for hospitalized patients; (2) administering candidate items to a calibration cohort; (3) estimating multidimensional item response theory models; (4) calibrating the item banks; (5) evaluating potential multidimensional computerized adaptive testing (MCAT) enhancements; (6) parameterizing the MCAT; (7) administering it to patients in a validation cohort; and (8) estimating its predictive and psychometric characteristics.

**Setting:** A large (2000-bed) Midwestern Medical Center.

**Participants:** The overall sample included 4495 adults (2341 in a calibration cohort, 2154 in a validation cohort) who were admitted either to medical services with at least 1 chronic condition or to surgical/medical services if they required readmission after a hospitalization for surgery (N=4495).

**Intervention:** Not applicable.

**Main Outcome Measures:** Not applicable.

**Results:** The FAMCAT is an instrument designed to permit the efficient, precise, low-burden, multidomain functional assessment of hospitalized patients. We tried to optimize the FAMCAT's efficiency and precision, as well as its ability to perform multiple assessments during a hospital stay, by applying cutting edge methods such as the adaptive measure of change (AMC), differential item functioning computerized adaptive testing, and integration of collateral test-taking information, particularly item response times. Evaluation of these candidate methods suggested that all may enhance MCAT performance, but none were integrated into initial MCAT parameterization.

**Conclusions:** The FAMCAT has the potential to address a longstanding need for structured, frequent, and accurate functional assessment among patients hospitalized with medical diagnoses and complications of surgery.

© 2021 The Authors. Published by Elsevier Inc. on behalf of American Congress of Rehabilitation Medicine. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Aging, frailty, and chronic disease account for more than 80% of United States health care spending, with the cost of care doubling for people with impaired mobility.<sup>1</sup> Increasing attention is being devoted to an important aspect of this serious problem: hospitalization rarely addresses and often accelerates the progressive functional losses of these groups.<sup>2-6</sup> Most importantly, a majority recover slowly, if at all, from hospital acquired functional losses and are consequently placed at a markedly increased risk of falls, institutionalization, rehospitalization, and even death.<sup>7-11</sup> Tellingly, these losses have contributed to a more than doubling of postacute care (PAC) spending in the past decade.<sup>12</sup>

Hospital-based rehabilitation has been proven to slow or prevent these losses, but its provision has been limited by human resource constraints and challenges in providing the right services to the right patient. In fact, a minority of patients who could benefit from rehabilitation services actually receive them. For example, many patients referred for physical therapy are never seen, and with the exception of specialized populations (eg, stroke, spinal cord injury, hip fractures), extended delays in treatment are common during which patients often remain bed-based.<sup>13,14</sup> Nurses are generally expected to mobilize patients who are not seen by

therapists; however, nurses confront formidable competing demands, and even ambulatory older patients spend the majority of their time in bed.<sup>15</sup>

These delays and omissions can be catastrophic. Up to 63% of older patients rapidly lose muscle mass<sup>16,17</sup> and decline in their mobility and capacity for self-care during even brief hospitalizations.<sup>8,18,19</sup> The expectation that they will regain this lost function is frequently not met, leading to institutionalizations and increased caregiver demands. Such outcomes are often avoidable because rehabilitation has been clearly shown to reduce care utilization, hospital lengths of stay, and PAC use in chronic diseases ranging from heart failure to cancer.<sup>20-22</sup>

An absence of a data-driven, standardized means to determine patients' rehabilitative needs is a critical barrier to preserving their function.<sup>10</sup> A new, more effective model is needed; however, increasing the demands on oversubscribed nurses and/or boosting therapist staffing are unlikely to be effective or scalable solutions. The use of mobility technicians or personal care assistants to implement simple but effective<sup>23</sup> mobility preservation care plans may offer promise. However, this approach requires a systematic and accurate means of matching of patients' needs with ability-

matched care plans. Historically, we have relied on clinician assessment as the sole basis for such matching, even though patient-reported information has shown value as a means of distinguishing inpatient care needs.<sup>24</sup> Human resource-intensive triage has proven a damaging bottleneck to timely service provision. This limitation has proven particularly pernicious for patients admitted with medical diagnoses because they are often frail with multiple comorbid conditions and are uniquely vulnerable to the disabling effects of even brief immobility.

The use of routinized functional measurements has been shown to be a feasible and scalable means of identifying patients' rehabilitation service needs. Specifically, the 6-clicks short forms (SFs) have shown promise as a timely means of identifying hospitalized patients who require therapy.<sup>25</sup> Although originally developed as a patient-reported outcome measure (PROM) for use in PAC, the 6-clicks is principally administered by nurses or therapists as a clinician-rated measure in acute care settings. When used in this manner among populations with orthopedic and neurologic conditions, 6-click scores associate with hospital discharge location and have proven useful for allocating therapy resource and discharge planning.<sup>26-28</sup> The 6-clicks has been less studied among patients hospitalized with medical diagnoses. In contrast, functional items from the Braden Scale for predicting pressure ulcer risk have been shown to be strongly associated with discharge destination in medical populations.<sup>29,30</sup> However, similar to the 6-clicks, the Braden Scale is provider administered with associated burdens and barriers. Moreover, neither tool has been scrutinized as a means of monitoring functional change over time.

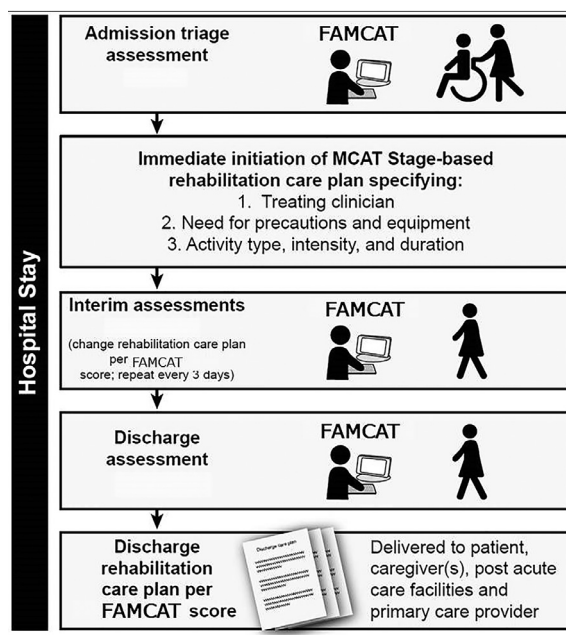
The R01-funded Computerized Adaptive Testing to Improve Delivery of Function-Directed Care project was designed to address the need for an easy-to-use, low burden, functional assessment tool with high discrimination applicable for patients hospitalized with medical diagnoses. In brief, the project sought to (1) develop a patient-reported, multidomain functional assessment tool focused on medically ill patients in acute care settings; (2) rigorously characterize the measure's psychometric performance; and (3) establish clinically actionable score strata for functional domains that would link directly to easily implemented mobility preservation plans irrespective of a patient's status.

This article provides a high-level overview of the multi-step process that our team pursued to realize these goals. Described below is the approach we used to develop the Functional Assessment in Acute Care Multidimensional Computer Adaptive Test (FAMCAT), a multidimensional item response theory (MIRT)-based measure of key functional domains among hospitalized patients admitted to medical services.

## Approach

### Overview, setting, and population

The FAMCAT was conceptualized as a means to guide patients, their caregivers, and inpatient and primary care providers in a continuing program of needs-matched function-directed activities during and after a hospital stay



**Fig 1** Anticipated integration of FAMCAT testing during and following a typical hospital stay.

(fig 1). Development of the FAMCAT was conceived as a multi-step process including the steps illustrated in figure 2: (1) expand and refine existing item banks to optimize salience for hospitalized patients; (2) administer candidate items to patients in the calibration cohort; (3) estimate MIRT models, calibrate item banks, and evaluate potential multidimensional computerized adaptive testing (MCAT) enhancements; (4) parameterize FAMCAT; (5) administer the FAMCAT to patients in validation cohort; and (6) estimate FAMCAT predictive and psychometric characteristics. Because the provision of rehabilitation services is more frequently inconsistent, delayed, and/or absent among patients on medical services or readmitted to surgical services, these subgroups comprised the FAMCAT target population. All research activities were conducted within the Mayo Clinic hospitals, Rochester, Minnesota, and were approved by the Mayo Clinic Institutional Review Board.

## Justification for defining project characteristics

### Rationale for using the extant Activity Measure of Post-Acute Care banks

Rather than develop all items de novo, we elected to use the Activity Measure of Post-Acute Care (AM-PAC) item banks<sup>31</sup> as a starting point. The item response theory (IRT)-modeled AM-PAC was the first multidomain functional PROM with the capability to direct care.<sup>31</sup> Its 3 domains, Mobility (131 items), Daily Activities (88 items), and Applied Cognitive (50 items), were established through factor, modified parallel, and Rasch analysis and encompass the dimensions of function essential for independence using data collected from patients in PAC settings.<sup>32,33</sup> One-third of the 1041-patient

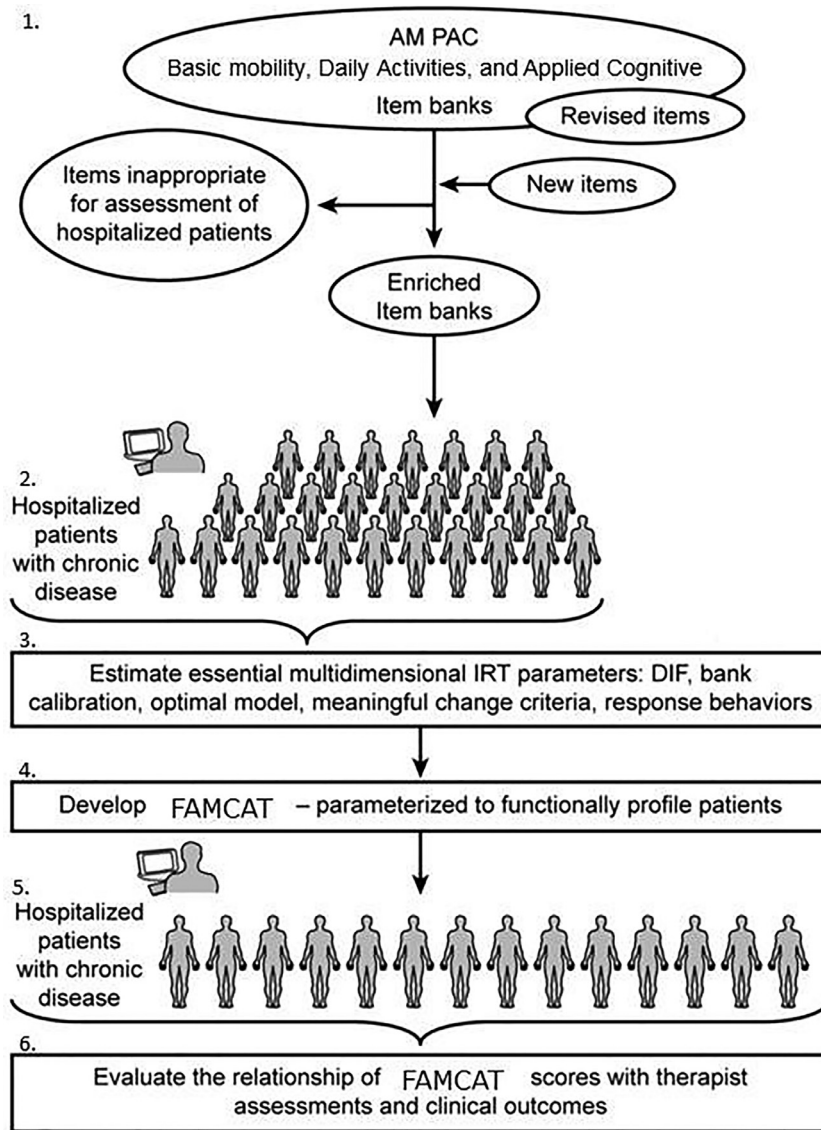


Fig 2 Sequential steps in FAMCAT development and testing.

cohort used to initially calibrate the AM-PAC item banks had complex or chronic medical conditions. Additionally, the AM-PAC banks were developed to align with the domains of the World Health Organization's International Classification of Functioning, Disability, and Health and therefore conform to a widely accepted conceptual framework. Moreover, extensive work has established the enhanced precision, reduced ceiling/floor effects and lessened respondent burden achieved when the AM-PAC domains were administered using a computerized adaptive testing (CAT) platform.<sup>34,35</sup> Importantly, the AM-PAC CAT's responsiveness in longitudinal monitoring of symptomatic and chronically ill patients has already been explored by members of our team.<sup>36</sup>

### Rationale for MCAT

A key project goal was to render repeated comprehensive yet precise functional assessments feasible within busy hospital settings where only a limited number of items can be administered. MIRT and MCAT allow for the simultaneous,

and hence more efficient, estimation of correlated traits.<sup>37-39</sup> Because the 3 AM-PAC domains are moderately correlated, MCAT administration offered a potential means of reducing the number of items required to achieve sufficient precision to inform clinical care.<sup>40</sup> MIRT models can be used to specify MCAT algorithms for item selection, although there is limited precedent for this approach in the medical field.

Although MIRT concepts have been available for many years,<sup>41,42</sup> only recently have computing capacity and estimation algorithms reached a level to permit realistic implementation.<sup>37,38,43-45</sup> The administration of MIRT-modeled item banks with an MCAT platform offers an opportunity to further enhance measurement efficiency because the MCAT, rather than selecting an item for a single scale at each stage of a CAT, selects an item that simultaneously provides the most information about the examinee's levels on *all* functional ability domains being assessed.<sup>39,44</sup> As such, an MCAT rapidly yields more precise score estimates with less respondent burden than would a series of unidimensional CATs.



## Rationale for enhancements to the MCAT

Implementation of MCAT alone was expected to result in significant gains in measurement precision and efficiency; however, additional enhancements to the MCAT algorithm were adopted as means of achieving even greater improvement. The project considered 3 enhancements for which there was a strong theoretical and anecdotal foundation. First, we proposed a novel strategy to address differential item functioning (DIF) that can occur when the probability of item responses varies across groups defined by age, education, ethnicity, and so on.<sup>46</sup> This means that on average, individuals from different subgroups but with the same level of functional ability may answer the item differently. Put another way, reporting difficulty with walking should depend only on the level of ability or disability in mobility and not on membership in a group, for example, male or female. Identifying the presence and magnitude of DIF in clinically integrated PROMs is essential to eliminating bias and addressing health care disparities.<sup>47</sup> A customary approach is to eliminate items that display DIF.<sup>48</sup> However, highly discriminating items may be lost in this way and, potentially, a different bias introduced—an inability to estimate traits with equal precision across subgroups.<sup>49</sup> CAT offers an alternate approach that we termed DIF-CAT.<sup>50</sup> In DIF-CAT, DIF information is incorporated into the MCAT item selection algorithm such that subgroup specific item parameters can be used for items that display DIF, provided that the MCAT was informed of a patients' subgroup membership before starting the test. DIF information was also used to lower the exposure rate (the frequency with which items are administered in a CAT) of items that displayed DIF.

Second, we proposed to leverage collateral test-taking information to enhance MCAT efficiency. More specifically, the amount of time that test takers require to respond to an item may provide information that can accelerate trait estimation. We hypothesized that longer response times may correlate with lower Applied Cognitive function estimates and that these data could be included in MIRT models to enhance precision. Hierarchical approaches model participants' responses and response times simultaneously.<sup>51</sup> These models have been used in academic assessments to identify cheating behaviors and to reduce the number of items required for trait estimation. Because AM-PAC response times vary substantially,<sup>52</sup> it is reasonable to test these models as means to enhance the efficiency of MIRT trait estimation.<sup>53,54</sup>

Last, we proposed to determine whether an Adaptive Measurement of Change (AMC)<sup>55,56</sup> approach could reduce the number of items administered on repeat assessments. In AMC, a CAT is administered at 2 (or more) time points. In practice, (1) the examinee's trait theta ( $\theta$ ) level from time 1 is used to begin the time 2 CAT, and (2) termination of the time 2 CAT occurs when sufficient evidence has been obtained to determine whether a statistically significant change has occurred. Thus, the AMC limits the second CAT to the minimal number of items needed to determine whether a respondent has changed from the time of the previous CAT session. This approach may substantially reduce respondent burden during repeat assessments, a highly desirable attribute in clinical assessment.

The project proposed to extend the AMC procedure to polytomous scored items based on the IRT models used in

MCAT and to extend the methodologies of AMC to multiple occasions of measurement to detect transitions between MCAT-defined mobility strata. This ability to detect transitions was thought to be clinically desirable because the functional status of medically ill patients may be highly dynamic, particularly after transitions to and from intensive care units, with important management implications.

## Item bank enrichment

A total of 44 AM-PAC items were deleted from the AM-PAC banks for lack of relevance to hospital settings, and 101 new items were added, yielding a total of 326 items across 3 domains: Basic Mobility (111 items), Daily Activities (108 items), and Applied Cognitive (107 items). Table 1 summarizes the enrichment of the AM-PAC candidate items in an effort to enhance their salience to hospitalized patients.

## Item bank culling

To adapt the item banks' coverage and content for hospitalized patients, panels of 8-9 clinical content experts were assembled for each AM-PAC domain. Because the AM-PAC banks were initially developed to assess patients in PAC settings, multiple items queried respondents about the degree of difficulty they experienced when performing activities with the gait aids and wheelchairs commonly used in those settings. Consensus was reached among the expert panel to remove these items because fewer patients use gait aids in the hospital, patients may not have their aids in the hospital, and inquiring about gait aids would increase the response burden.

## Expansion

Subdomains were identified within each domain, and the experts assigned the retained AM-PAC items to the domain and subdomains. Some items were reassigned from their original AM-PAC domain to a different domain by the experts; for example, "How much difficulty do you currently have operating an ATM to get cash or make deposits?" was moved from Daily Activities to Applied Cognitive. The experts identified content gaps in subdomain coverage across the entire range of each trait and provided potential sources of extant items to fill the deficits. In addition to legacy instruments suggested by the expert panels, the IRT-modeled Patient-Reported Outcomes Measurement Information System (PROMIS) and Quality of Life in Neurological Disorders banks were reviewed.<sup>57</sup> Items selected from these sources were edited to conform to the stem structure and response options of the AM-PAC items. Most items began with "How much DIFFICULTY do you currently have..." and presented response options "unable," "a lot," "a little," and "none"; a small percentage of items began with "How much HELP from another person do you currently need..." and used response options "total," "a lot," "a little," and "none." Persistent coverage deficits were addressed by writing new items related to the limited activities that can be performed in a standard hospital room. A total of 43 de novo items were generated and tested with inpatients to confirm

**Table 1** FAMCAT item bank expansion summary

Domain	Subdomain	No. of Original AM-PAC Items Retained (n)	No. of Items Added/ Modified From Extant Sources (n)	No. of Items Written De Novo By Study Team (n)	Total No. of Items In Initial Calibration Cohort (n)	No. of Linking Items (n)
Mobility	Ambulation	15	6	3	24	4
	Carrying/reaching	11	0	8	19	0
	Changing body position	9	0	2	11	0
	Maintaining body position	7	2	2	11	0
	Stair climbing	15	0	0	15	4
	Transfers	19	0	0	19	0
	Other	12	0	0	12	0
	Total for Mobility domain	88	8	15	111	8
Daily Activities	ADL	26	3	3	32	2
	Appendicular strength	14	3	5	22	1
	Dexterity	25	1	3	29	4
	IADL	13	0	2	15	1
	Reaching	8	0	2	10	0
	Total for Daily Activities domain	86	7	15	108	8
Applied Cognitive	Communication: verbal	13	8	1	22	3
	Communication: written	7	3	0	10	1
	Decision making	1	3	0	4	0
	Environmental awareness	1	0	1	2	0
	Problem solving/executive functioning	14	11	1	26	3
	Procedural memory	2	3	0	5	0
	Processing speed	1	4	1	6	0
	Social awareness	3	0	0	3	0
	Understanding instructions	4	1	6	11	0
	Working memory	5	10	3	18	1
	Total for Applied Cognitive domain	51	43	13	107	8

Abbreviations: ADL, activities of daily living; IADL, instrumental activities of daily living.

understanding.<sup>58</sup> A series of 6 teleconferences were held throughout the item bank expansion process to allow the expert panel to reach consensus on recommendations and finalization of the item bank.

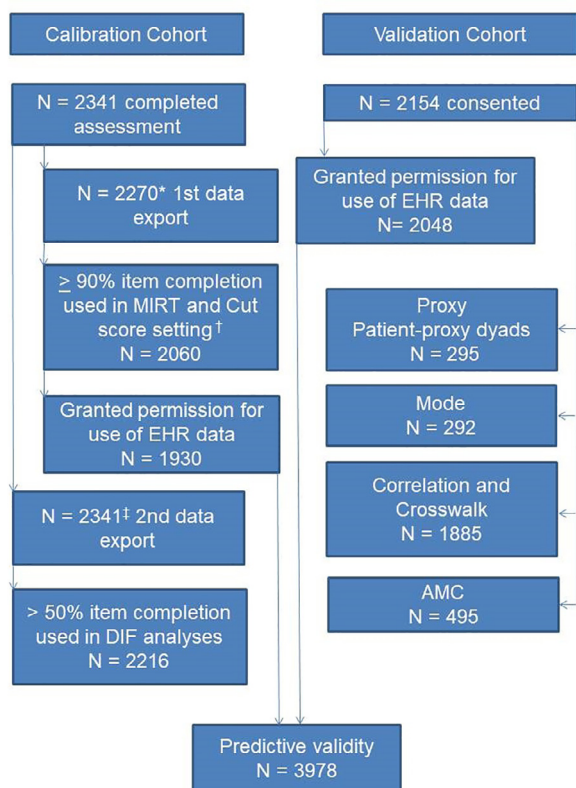
### Calibration cohort enrollment

Participants (n=2341) were recruited from the Mayo Clinic Hospital and identified through a well-established electronic search tool. Minority recruitment was enhanced using the search tool to optimize demographic representation. During the 13-month initial data collection interval (May 2016-June 2017) the tool was used to identify patients admitted to inpatient medical services over the preceding 24 hours with at least 1 chronic condition. Although the study's primary focus was patients admitted to medical services, patients with complicated postoperative courses were also considered appropriate and approached if they required readmission after a hospitalization for surgery. Figure 3 outlines the flow of participants and their data through the calibration and validation cohort studies.

Patients' electronic medical records were reviewed to determine eligibility: no requirement for ventilatory support

other than continuous positive airway pressure or intermittent bilevel positive airway pressure; no use of cognitive depressant medications apart from soporifics, antipsychotics, anxiolytics, analgesics, or antidepressants; ability to respond to orally administered questions; and fluency in English adequate to respond to the items. The Mini-Cog<sup>35</sup> was collected for use as a covariate in the analysis of the Applied Cognitive domain. Patients were interviewed on a single occasion immediately after providing written informed consent and Health Insurance Portability and Accountability Act of 1996 (HIPAA) authorization.

Purposive sampling was used to ensure adequate representation of demographically and clinically defined subgroups spanning the entire trait range. Table 2 lists the demographic and clinical characteristics of the calibration and validation cohorts. Once the pool of potentially eligible patients was established on a given day, targeted recruitment was used to ensure that the sample maintained adequate subgroup representation for DIF analyses with the following characteristics: (1) roughly equal numbers in each age stratum (<60, 60-75, and >75 years); (2) ≥15% high school noncompletion (comparable with national levels); and (3) ≥15% with moderate to severe pain.<sup>59-61</sup> Additionally, recruitment efforts were coordinated across hospital



**Fig 3** Participant flow diagram for calibration and validation cohorts. \*An initial batch 1 data export was performed after 500 participants had been assessed to identify linking items. The identification of linking items prior to completing batch 1 data collection allowed a seamless transition from batch 1 to batch 2 collection because batch 2 included the linking items. Data from this initial pull were used for the MIRT models. †Responses were retained from calibration cohort members who answered at least 90% of the administered items. ‡The complete calibration cohort data set was used for the DIF analyses. These data differed in that they included the batch 1 data collected following the initial export.

services (ie, cardiac, gastrointestinal, organ transplant, pulmonary, medical oncology, general internal medicine, etc) and hospital floors/buildings to ensure a clinically diverse sample. Patients admitted to neurology services or readmitted to neurosurgical services were not recruited because therapy is routinely provided to these patients, and they are consequently at a lesser risk of preventable hospital acquired disablement.

## Calibration data collection

### Item batching

A key goal of the FAMCAT project was to longitudinally assess patients' risk for hospital-acquired disability due to immobility. This subpopulation was thought to be best represented among patients admitted to medical services and those readmitted for complications of surgical procedures.

Given the frequently stressed, symptomatic, and ill status of these patients, answering all 326 candidate items was deemed neither humane nor practical. The items were therefore separated into 4 batches of roughly equal size. To create batches with equal domain, subdomain, and trait level representation, the IRT item information characteristics were obtained, when available, and items were positioned along each trait continuum. Four representative batches were manually created with checks to assure subdomain representation. Because it was critical that high-quality, DIF-free linking items be selected from the first batch, this batch was slightly larger,  $n=110$ . Twenty-four linking items (8 per domain) were identified in the first batch based on maximizing the information coverage along the wide range of the trait levels (ie, standardized trait levels from  $-3$  to  $3$ ) for each domain and were included in batches 2-4. Minus the 24 linking items, which were common to all batches, the batch sizes were 86, 72, 73, and 71 items. The 4 batches were programmed into the Qualtrics survey administration and storage platform.<sup>a</sup>

### Item administration

Research assistants read items from each batch to participants from the Qualtrics interface and were instructed not to interpret items or offer other guidance. Items within batches were organized into blocks according to domain. The order of blocks within batches and the order of items within blocks were randomized. Participants had the option to change their answers until the research administrator advanced to the next question. Once  $>500$  participants responded to the items in first batch, the 24 high-performing, DIF-free linking items noted above were identified. The linking items were added to batches 2-4. A sample of  $n>500$  was targeted for each batch, with an anticipated incompleteness rate of 10%. The final number of respondents for batches 1, 2, 3, and 4 were 701, 542, 555, and 543, respectively, as outlined in Table 3. Participants in the calibration data collection had a mean age of 61.8 years, 54% were male, 96% were non-Hispanic white, and 78% had 2 or more comorbidities.

### Electronic health record abstraction

Participants' demographic and clinical information was electronically abstracted from the Mayo Clinic Unified Data Platform, which stores aggregated clinical and administrative data. In addition to comorbidities assigned in the 12 months prior to discharge, discharge location, 30-hospital readmission status, admission/discharge diagnosis, functional items from the Braden Scale recorded by nurses, and FIM items recorded by therapists were abstracted. The 2 functional components of the Braden Scale (ordinal assessments of the degree of physical activity and the ability to change/control body position) for predicting pressure ulcer risk are charted by nurses on every patient at least twice daily.<sup>62</sup> Several items from the FIM, including those related to supine-to-sit and sit-to-stand transfers, ambulation, dressing, and toileting, were abstracted for all participants who underwent therapy evaluations during their hospitalizations.<sup>63-66</sup>

**Table 2** Demographic and clinical characteristics of the FAMCAT validation and calibration cohorts

Characteristics	Validation Cohort, n=2050	Calibration Cohort, n=2024
Age (y)		
mean $\pm$ SD	61.4 $\pm$ 16.0	63.6 $\pm$ 16.0
median (IQR)	63.0 (52.0-72.0)	66.0 (55.0-75.0)
Sex, n (%)		
Female	952 (46.4)	933 (46.1)
Male	1098 (53.6)	1091 (53.9)
Charlson Comorbidity Index		
Charlson		
mean $\pm$ SD	1.3 $\pm$ 1.4	1.2 $\pm$ 1.4
median (IQR)	1.0 (0-2.0)	1.0 (0-2.0)
Charlson Severity		
mean $\pm$ SD	2.3 $\pm$ 2.6	1.8 $\pm$ 2.4
median (IQR)	2.0 (0-3.0)	1.0 (0-3.0)
Charlson Severity and Age		
mean $\pm$ SD	4.1 $\pm$ 3.1	3.8 $\pm$ 2.9
median (IQR)	4.0 (2.0-6.0)	3.0 (2.0-5.0)
Hospital length of stay (d)		
mean $\pm$ SD	7.1 $\pm$ 8.2	4.4 $\pm$ 5.3
median (IQR)	5.0 (3.0-8.0)	3.0 (2.0-5.0)
Discharge location, n (%)		
Home with/without home care	1822 (89.2)	1868 (93.0)
Intensive inpatient rehabilitation or skilled Nursing facility	221 (10.8)	140 (7.0)
Missing	7	16
PT consultation, n (%)		
	300 (14.6)	111 (5.5)
OT consultation, n (%)		
	236 (11.5)	81 (4.0)
30-d Hospital readmission, n (%)		
	103 (5.3)	80 (4.5)
Missing	118	252
Admission diagnosis, CCS category, n (%)		
Diseases of the blood and blood-forming organs and immune system disorders	41 (2.0)	31 (1.5)
Diseases of the circulatory system	268 (13.2)	684 (33.9)
Diseases of the digestive system	369 (18.2)	296 (14.6)
Endocrine, nutritional, and metabolic disease	64 (3.1)	86 (4.3)
Diseases of the genitourinary system	134 (6.6)	92 (4.5)
Infectious and parasitic diseases	164 (8.1)	109 (5.4)
Injury, poisoning, and certain other consequences of external causes	137 (6.7)	113 (5.6)
Mental, behavioral, and neurodevelopmental disorders	10 (0.5)	10 (0.5)
Diseases of the musculoskeletal system and connective tissue	69 (3.4)	51 (2.5)
Neoplasms	492 (24.2)	198 (9.8)
Diseases of the nervous system	26 (1.3)	22 (1.1)
Diseases of the respiratory system	148 (7.3)	160 (7.9)
Symptoms, signs, and abnormal clinical/laboratory findings	56 (2.8)	88 (4.4)
Other*	52 (2.56)	80 (4.0)

Abbreviations: CCS, chronic condition software; OT, occupational therapy; PT, physical therapy.

\* Other includes 5 CCS categories: diseases of the ear and mastoid process; diseases of the eye and adnexa; congenital malformations, deformations, and chromosomal abnormalities; pregnancy, childbirth, and the puerperium; and diseases of the skin and subcutaneous tissue.

Those who died or were transitioned to hospice were excluded; these statistics are calculated using the cohort data from the prediction article.

## MIRT modeling

We conducted an exploratory item factor analysis on each batch separately and, subsequently, on the combined data. Models with 1-, 2-, 3-, and 4-factor structures were compared. Relative model fit indices, Akaike's information

criterion<sup>67</sup> and the Bayesian information criterion,<sup>68</sup> revealed that a 3-factor model outperformed the 1- and 2-factor models consistently, but a 4-factor model seemed to fit the data the best. The 4-factor structure suggested dividing the factor of Applied Cognitive into 2 additional factors. However, because probing the underlying factor structure of



**Table 3** Item bank completion by batch

Batch	No. of Items in Batch*	Patients Accrued (n)	Patients Completed All Items in the Batch, n (%)	Patients Completed at Least 1 Item, n (%)
1	110	701	481 (68.6)	698 (99.6)
2	96	542	261 (48.2)	536 (98.9)
3	96	555	291 (52.4)	547 (98.6)
4	96	543	351 (64.6)	541 (99.6)
Total		2341	1384 (59.1)	2322 (99.2)

\* Includes 24 linking items that are common to all batches.

Applied Cognitive was beyond the focus of the current study, we decided to use a 3-factor IRT model. Then, we used the Expectation-Maximization algorithms implemented in *flexMIRT* for 3-factor multidimensional graded response model (MGRM) calibration.<sup>69</sup> All item parameters were properly recovered, and their standard errors were from 0.06-0.39, with an average of 0.24.

### Unidimensional and multidimensional DIF assessment

Hypotheses were generated<sup>49,70</sup> as per recommended best practices for DIF analyses,<sup>49,70-72</sup> on the basis of expert qualitative review regarding the likely presence and direction of DIF for all items with respect to age, race, sex, and duration of time in the hospital. In parallel with hypothesis generation, we examined dimensionality across groups as the first step in a hierarchy of invariance tests,<sup>73,74</sup> per the National Institutes of Health PROMIS guidelines as recommended by Reise et al.<sup>75,76</sup> Initial DIF estimates were obtained by treating each item as a “studied” item, while using the remainder as “anchor” (DIF-free) items. We used a modified “all-other” approach that included “iterative purification.” We then used a unidimensional DIF test, the IRT-Wald statistic contained in Item Response Theory for Patient Reported Outcomes,<sup>77</sup> to assess DIF in each of the item banks. Items showing DIF were excluded from the DIF-free anchor set at each iteration until no items showed DIF, and this set was used for final determination of DIF. A model was constructed with all parameters constrained to be equal across comparison groups for the anchor items and item parameters for all studied items freed to be estimated distinctly. An overall simultaneous joint test of differences in the discrimination (“*a*”) or severity (“*b*”) parameters was performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF was detected when the *b* parameters differed and nonuniform DIF when the *a* parameters differed. To assess DIF magnitude and effect, noncompensatory DIF,<sup>78</sup> reflecting group difference in expected item scores,<sup>79</sup> was used for DIF magnitude assessment; such effect size estimation has been recommended to identify salient DIF.<sup>80-85</sup> Summing the expected item scores provided differences in “test” response functions,<sup>86</sup> an index<sup>78,87</sup> of scale-level effect.

Initially, we planned to perform multidimensional DIF analyses. Unidimensional IRT DIF tests were used instead because recent simulation studies by members of our team demonstrated that sample size requirements for accurate

estimation of item parameters for the multidimensional model was at least 500.<sup>88</sup> The sample sizes of subgroups defined by race, sex, age, and duration of hospital stay were not large enough to perform multidimensional IRT DIF testing. A recent simulation study demonstrated that modeling DIF as unidimensional may be as accurate as multidimensional models for determining effect sizes for binary data.<sup>89</sup> Moreover, the initial dimensionality analyses for each domain examined supported a unidimensional approach to DIF detection within domains.

### Evaluation of collateral test-taking information

We assessed the utility of using participants’ response times to items using van der Linden’s hierarchical modeling framework.<sup>90</sup> At the measurement model level, the MGRM was used for modeling item responses, whereas a lognormal model was used to model item response times. At the higher-order model level, patients’ 3-dimensional latent traits and the unidimensional latent speed were correlated. Moreover, because during the field testing of the items an interviewer read each item to a patient and recorded their responses and response times, the interviewer was also included as a nominal covariate in the hierarchical model. The hierarchical model was fitted to all batches of data via a concurrent calibration. Results showed that adding response time information did not affect the item parameter estimates and their standard errors significantly. However, adding response time information helped reduce the standard error of patients’ multidimensional latent trait estimates, but adding interviewer as a covariate did not result in further improvement. Hence, using the MGRM for item parameter calibration is enough, but using response time as collateral information would help improve FAMCAT efficiency, although we ultimately chose not to incorporate response times in the MCAT algorithm.

### Defining clinically actionable FAMCAT score strata

There is ample precedent for using score strata from IRT-modeled PROMs to bin patients into clinically relevant and actionable categories.<sup>91,92</sup> Figure 4 illustrates the 4 clinically actionable levels that were hypothesized for each domain as well as their definitions. Three parallel strategies were used to identify candidate cut scores to delineate clinically relevant score strata in each domain using estimates


	Basic Mobility	Daily Activity	Cognition Communication
<p><b>Very low fall &amp; dependency risk</b></p> 	<p><i>Independent ambulator</i></p> <p>A. High-level community ambulator → Endurance trained</p> <p>B. Low → moderate level community ambulator</p>	<p><i>Independent with daily activities</i></p> <p>A. No difficulty with high level activities (shopping, housekeeping, yard work)</p> <p>B. Difficulty with high level, but no difficulty with low level activities (e.g. bathing, hygiene)</p>	<p><b>Independent communicating high level constructs and needs</b></p>
<p><b>Low to Intermediate fall &amp; dependency risk</b></p>	<p><i>Requires supervision or simple assistance</i></p> <p>A. Stand by assist → Contact guard, high fall &amp; dependency risk</p> <p>B. Assistive device needed (walker, EZ stand, etc.)</p>	<p><i>Requires supervision or simple assistance</i></p> <p>A. Requires supervision and cuing</p> <p>B. Requires assistive devices</p>	<p><b>Independent communicating low level constructs and needs</b></p>
<p><b>High fall &amp; dependency risk</b></p>	<p><i>Requires skilled assistance</i></p> <p>A. Requires minimal to moderate assistance with ambulation</p> <p>B. Maximal assistance out of bed, or bed-based but participatory</p>	<p><i>Requires skilled assistance</i></p> <p>A. Requires minimal to moderate assistance with low level activities</p> <p>B. Maximal assistance out of bed, or bed-based but participatory</p>	<p><b>Requires assistance to communicate low level constructs and needs</b></p>
<p><b>Bed-based</b></p>	<p>Non-participatory</p>	<p>Non-participatory</p>	<p><b>Unable to communicate low level constructs and needs</b></p>

Fig 4 Four hypothesized levels for each FAMCAT domain that inform individualized mobility preservation plans.

derived from the Functional Assessment for Acute Care Multidimensional (FAM) IRT models.

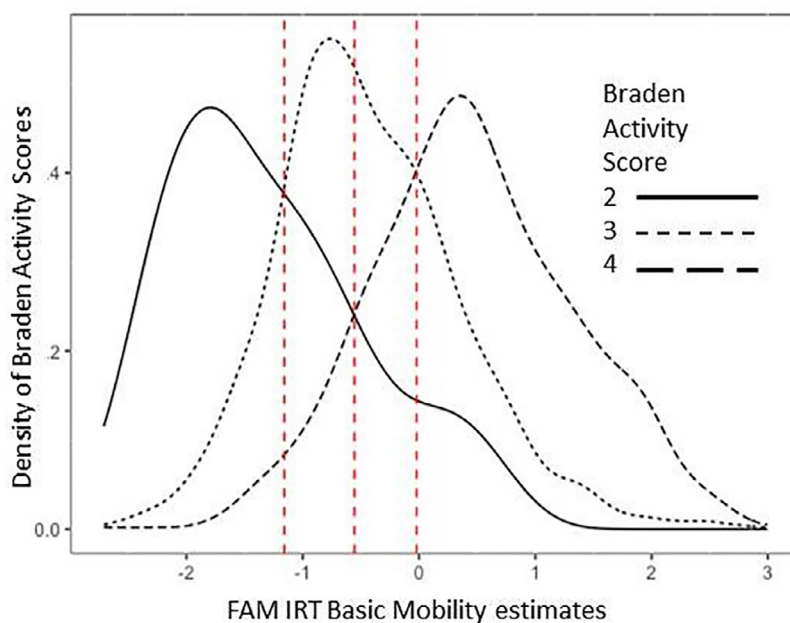
Although 4-level stratification, as depicted in figure 4, was initially anticipated for all domains, a single cut score for the Applied Cognitive domain was eventually adopted as being more clinically actionable. This cut score was conceptualized as a means of distinguishing patients with potentially severe enough cognitive impairment that their Basic Mobility and Daily Activity scores should be acted on cautiously because of a need for greater supervision or assistance than might be suggested by their mobility and activity scores alone.

**Graphic and statistical approaches to identify candidate cut scores**

Ordered categorical ratings representing constructs similar to those estimated by the FAM IRT Daily Activity and Basic Mobility models were available in electronic health record (EHR) data for the 2060 calibration cohort participants. These ratings were provided by nurses as well as physical and occupational therapists. However, physical therapy and occupational therapy assessments occurred for only 15% and

12% of participants, respectively. We considered these therapist assessment data insufficient to serve as the basis for establishing cut scores and therefore relied on Braden Activity Scores (BAS) entered by nurses. The Braden Activity Score is 1 of 6 subscales that comprise the Braden Scale,<sup>93</sup> which is used to predict pressure ulcers. The Braden Activity Score assesses mobility with a 4-point ordinal scale: 1 (“patient is confined to bed”); 2 (“severely limited or non-existent ability to walk; patient cannot bear his own weight and/or must be assisted into chair or wheelchair”); 3 (“patient walks occasionally during the day but for very short distances, with or without assistance; spends majority of each shift in bed or chair”); or 4 (“patient walks outside the room at least twice a day and inside the room at least once every 2 hours during waking hours”).

No participants were rated as “bed-based,” hence this category was “missing” from nurse ratings. To find cut scores along the Basic Mobility latent trait that maximized the consistency of classification decisions between FAM IRT estimates and nurse ratings, we plotted the smoothed frequency distribution of the FAM IRT Basic Mobility estimates for subgroups classified by nurse Braden ratings as depicted in figure 5. The 3 intersection points from the 3



**Fig 5** Smoothed frequency distributions of the basic mobility MIRT model estimates for subgroups classified by nurse mobility ratings.

distribution curves were considered as cut scores. On average, classifications based on the FAM IRT Basic Mobility predictions and Braden activity item agreed 69% of the time. We therefore plotted smoothed frequency distributions to establish candidate cut scores for the Daily Activity FAM IRT estimates as well, even though the Braden activity and the FAM Daily Activity items evaluate overlapping but distinct constructs. FAM IRT and Braden activity item classifications agreed 59.6% of the time for the Daily Activity domains. Mini-Cog scores were used to determine a single candidate cut score for the FAM IRT Applied Cognitive estimates. Mini-Cog and FAM IRT Applied Cognitive classifications agreed 71.8% of the time.

### Bookmark approach to identify candidate cut scores

We used a “bookmark” approach derived from educational settings whereby experts use a data-driven consensus process for setting standards for academic performance.<sup>94-96</sup> A panel of experts was convened composed of 3 rehabilitation physicians, 3 occupational therapists, and 3 physical therapists, all specialized in the care of medically ill hospitalized patients. The modified Delphi technique involved 3 rounds: independent cut score designation, feedback and summary of the independent cut score, and then finalization of the cut score with consensus.

### Subgroup analyses to identify candidate cut scores

Among the subgroup of participants dismissed from the hospital within 48 hours of testing, mean FAM IRT domain score differences were compared between patients who went to inpatient facilities, home with rehabilitation services, or home without services.

### Consensus

Final FAM IRT domain cut scores were established by a second expert consensus process. A panel that was distinct from participants in the bookmark approach, described above, and composed of 3 occupational therapists, 3 physical therapists, 3 physicians, and 3 nurses reviewed item maps with each of the candidate cut points established using the methods outlined above. The final consensus process considered the “bookmark”-derived cut points, those established through the graphic/statistical approach, as well as the effect of collapsing and/or subdividing stages. A modified Delphi process was used to determine the final cut points.

### FAMCAT algorithm development and testing

We developed MCAT algorithms for item selection during FAMCAT test sessions. These algorithms were evaluated through a series of Monte Carlo simulations for implementation in the FAMCAT.<sup>97</sup> Within the context of FAMCAT, different methods for selecting the next item to be administered were compared in Monte Carlo simulations to determine the most effective method for use in the FAMCAT. Java programmers programmed the final FAMCAT algorithms. Proper functioning of the FAMCAT algorithm, user interface, and storage and reporting aspects of the software were validated followed by FAMCAT beta testing and final software adjustments prior to FAMCAT release for data collection from the validation cohort.

### FAMCAT validation and psychometric assessment

The final FAMCAT algorithm was specified, tested, and programmed into the FastTest administration platform.<sup>b</sup>

Convergent and predictive validity as well as the presence and magnitude of proxy and mode effects were comprehensively assessed using data collected from a validation cohort composed of 2154 hospitalized patients who contributed a total of 2887 assessments, as outlined in figure 3.

## Study designs

The predictive validity study used a prospective design to estimate correlations between participants' FAMCAT scores and downstream events: (1) discharge to home or PAC and (2) 30-day readmission, which were electronically abstracted from the EHR. The convergent validity study used a cross-sectional design to estimate correlations between FAMCAT scores and patient- and clinician-rated functional outcomes collected concurrently with the FAMCAT. A mode study used a randomized cross-sectional design to assess the presence and magnitude of mode effects between interview and tablet-based FAMCAT administration. Last, a proxy study used a cross-sectional design to estimate the presence and magnitude of proxy effects when the FAMCAT was administered using tablets to 295 patient-proxy dyads.

## Participants

### Patients

The recruitment strategies used for the FAMCAT validation and psychometric assessments of the validation cohort ( $n=2154$ ) were identical for all studies and were similar to those used for the calibration phase. However, the Mayo Clinic transitioned to the Epic EHR in the interval between the enrollment of the calibration and validation cohorts. Therefore, for the validation cohort an Epic Report,<sup>c</sup> rather than the search of the administrative Mayo Clinic Unified Data Platform, was run daily to identify potentially eligible patients. EHR problem lists were reviewed to remove patients with combative behavior, active drug and/or alcohol withdrawal, and advanced dementia. Potential participants' nurses were queried regarding additional eligibility criteria: English fluency, sufficient auditory acuity to respond to the items, and no receipt of sedation within the past 6 hours. Once a patient's nurse cleared them for participation, the study was described to the patient. Receptive patients provided informed consent and signed a HIPAA authorization form. The majority of participants provided data at only 1 time point; however, given our aim of estimating the AMC, patients were administered the MCAT up to 4 times during their hospital admission. If patients were readmitted to the hospital, they were eligible to participate in additional FAMCAT sessions. Of 2887 FAMCAT patient assessments, 885 were follow-ups.

### Clinicians

We collected data to assess the FAMCAT's convergent validity from nurses and physical therapists. Nurses caring for participants provided informed consent on 1 occasion. Because no personal health information was collected from nurses, they were not required to sign HIPAA authorizations. Because therapists provided data in the EHR in the course of delivering routine clinical care, they did not provide informed consent.

### Proxies

To be eligible to participate in the proxy study, proxies were required to have resided with the patient for a minimum of 1 week and to have last resided with them no more than 2 days prior to admission. Proxies provided oral informed consent; however, because personal health information was not collected from them, they were not required to provide HIPAA authorization.

## Data collection procedures

Data were collected by research coordinators in the participants' hospital rooms between 7 AM and 5 PM on weekdays. Two approaches were used. For FAMCAT sessions administered by interview, items were read to participants who communicated their responses orally. Alternatively, for sessions administered via tablet, iPads were used for the FAMCAT items.<sup>d</sup> PROMIS Physical Function (PF) items were administered orally to all participants. Participants were given as much time as they needed for tablet sessions. Irrespective of administration mode and similar to data collection during the validation study, research coordinators were instructed not to interpret or explain the items during testing sessions.

For all sessions, the FAMCAT was first administered to participants followed by the PROMIS PF SF items. Data were collected from patients' nurses for use in the convergent validity study either immediately prior to or after participants' FAMCAT sessions. Nurses were not present in participants' rooms during FAMCAT administration.

Person-reported outcomes, in addition to the FAMCAT, and data automatically abstracted from the EHR for each psychometric study were as follows:

## Measures

### 6-clicks

This AM-PAC SF instrument has 6 questions evaluating a person's need for assistance in completing distinct functional mobility activities.<sup>25,98</sup> Based on clinician judgment, each question is scored on a 4-point ordinal scale, where a score of 1 indicates that the person is unable to complete the task and 4 indicates that the person is independent in completing that activity.

### Johns Hopkins—Highest Level of Mobility

The Johns Hopkins—Highest Level of Mobility evaluates general mobility over a fixed observation period.<sup>99</sup> Scoring is based on a person's observed activity as a 1-item scale with 8 ordinal response options: 1=only lying, 2=bed activities, 3=sitting at edge of bed, 4=transferring to chair, 5=standing for  $\geq 1$  minute, 6=walking  $\geq 10$  steps, 7=walking approximately  $\geq 7.5$  m ( $\geq 25$  ft), and 8=walking approximately  $\geq 75$  m ( $\geq 250$  ft).<sup>100</sup>

### Braden Activity Score

The Braden Activity Score was described previously.

### Eight-item PROMIS PF SF

The PROMIS PF validated 8-item SF assesses mobility and daily living activities.<sup>101-103</sup> Because PROMIS items are not



scored as sums but rather on a standardized T score metric using IRT, scores obtained from different item subsets are readily comparable.

## Analyses

### Predictive validity study

To assess the FAMCAT's predictive validity, associations between FAMCAT scores and participants' discharge locations were estimated: home, home with rehabilitation services, skilled nursing facility, inpatient rehabilitation facility, and long-term acute care hospital. In addition, 30-day hospital readmissions were ascertained. We compared the FAMCAT's capacity to predict discharge location with the 6-clicks and PROMIS PF SF.

### Convergent validity study

We characterized the FAMCAT's convergent validity by estimating correlations of FAMCAT scores with clinician-rated Johns Hopkins–Highest Level of Mobility, 6-clicks, and Braden Activity Score and self-rated PROMIS PF SF functional outcomes.

### Mode study

A 3-way multivariate analysis of variance was performed to determine whether test mode as well as the patients' sex and age were associated with at least 1 of the 3 latent traits: Applied Cognitive, Daily Activity, and Mobility.

### Proxy study

To determine if FAMCAT scores (ie,  $\theta$  estimates) from the proxies were significantly different from those obtained from the patients, a repeated measures multivariate analysis of variance was conducted with the independent variables being the sex and the age of the patient, as well as the patient vs proxy variable. Additional analyses directly compared each patient's ratings with those of their proxies.

## Discussion

The FAMCAT was developed to permit the efficient, low-burden, and precise functional assessment of patients admitted to medical services or readmitted to surgical services for postoperative complications. FAMCAT development was guided by the need to balance 3 key requirements: (1) efficiency, to permit integration into busy clinical work flows; (2) absence of clinical burden because oversubscribed clinicians have proven to be limited in their ability to consistently record high-quality function data; and (3) precision, for the timely, accurate individualization of patients' mobility preservation plans. We endeavored to further optimize the FAMCAT's efficiency by applying cutting edge methods that have gained traction in academic assessment but have yet to be used in clinical contexts, namely the use of collateral test-taking information (response times) and the AMC.

Greater reliance on PROM-based functional assessment among hospitalized patients offers several significant advantages. Principal among these is the capability of performing frequent reassessments without burdening clinicians. Such frequency is critical to detect the rapid changes that often mark the functional status of patients in acute care. These

individuals frequently transfer in and out of intensive care units; experience abrupt restoration of homeostasis; and/or respond to treatments that eliminate ischemia, infection, and inflammation. Moribund patients incapable of independent mobility may be transformed in a matter of days. Such changes have clear and immediate implications for patients' mobility requirements and precautions as well as their PAC needs. The means to detect clinically actionable changes in a precise and timely manner currently eludes the capabilities of most health care systems. Without the development and implementation of better inpatient assessment systems, function-directed care will remain a haphazard iteration of a more effective, needs-matched future state.

## Conclusions

The effort to develop the FAMCAT used both novel and established methods to address the long-standing need for a way to obtain frequent, structured, sensitive, and accurate functional assessments of hospitalized patients without increasing clinician workloads. Whether or not this instrument can achieve its goal is currently under assessment with the first assessments of these efforts scheduled to appear in a 2021 supplement of the *Archives of Physical Medicine and Rehabilitation*.

## Suppliers

- Qualtrics; Qualtrics International.
- FastTest; ASC.
- Epic Report; Epic Systems Corporation.
- iPad; Apple Inc.

## Corresponding author

Jeffrey R. Basford, MD, PhD, Mayo Clinic College of Medicine, 200 Second Street SW, Rochester, MN 55905 *E-mail address:* [basford.jeffrey@mayo.edu](mailto:basford.jeffrey@mayo.edu).

## Acknowledgment

We thank Alan Jette, PT, PhD, whose work and insights served as a major basis of our efforts.

## References

- Robert Wood Johnson Foundation. Chronic care: making the case for ongoing care. Available at: <http://www.rwjf.org/content/dam/farm/reports/reports/2010/rwjf54583>. Accessed May 11, 2013.
- Shi L, Hospitals Singh D. Essentials of the US health care system. Sudbury: Jones and Bartlett Publishers; 2005. p. 173-96.
- Rosenberg C. Imposing a new order: sources of change. The Care of strangers: the rise of America's hospital system. New York: Basic Books Inc Publishers; 1987. p. 288-97.
- Davydow DS, Hough CL, Levine DA, Langa KM, Iwashyna TJ. Functional disability, cognitive impairment, and depression after hospitalization for pneumonia. *Am J Med* 2013;126:615-24.

5. Sager MA, Franke T, Inouye SK, et al. Functional outcomes of acute medical illness and hospitalization in older persons. *Arch Intern Med* 1996;156:645-52.
6. Boyd CM, Ricks M, Fried LP, et al. Functional decline and recovery of activities of daily living in hospitalized, disabled older women: the Women's Health and Aging Study I. *J Am Geriatr Soc* 2009;57:1757-66.
7. Zisberg A, Shadmi E, Sinoff G, Gur-Yaish N, Srulovici E, Admi H. Low mobility during hospitalization and functional decline in older adults. *J Am Geriatr Soc* 2011;59:266-73.
8. Mahoney JE, Sager MA, Jalaluddin M. New walking dependence associated with hospitalization for acute medical illness: incidence and significance. *J Gerontol A Biol Sci Med Sci* 1998;53:M307-12.
9. Brown CJ, Friedkin RJ, Inouye SK. Prevalence and outcomes of low mobility in hospitalized older patients. *J Am Geriatr Soc* 2004;52:1263-70.
10. Ettinger WH. Can hospitalization-associated disability be prevented? *JAMA* 2011;306:1800-1.
11. Boyd CM, Landefeld CS, Counsell SR, et al. Recovery of activities of daily living in older adults after hospitalization for acute medical illness. *J Am Geriatr Soc* 2008;56:2171-9.
12. Miller ME. Medicare post-acute care reforms. Washington (DC): Medicare Payment Advisory Commission; 2013.
13. Grill E, Huber EO, Gloor-Juzi T, Stucki G. Intervention goals determine physical therapists' workload in the acute care setting. *Phys Ther* 2010;90:1468-78.
14. Pedersen MM, Bodilsen AC, Petersen J, et al. Twenty-four-hour mobility during acute hospitalization in older medical patients. *J Gerontol A Biol Sci Med Sci* 2013;68:331-7.
15. Brown CJ, Redden DT, Flood KL, Allman RM. The underrecognized epidemic of low mobility during hospitalization of older adults. *J Am Geriatr Soc* 2009;57:1660-5.
16. Evans WJ. Skeletal muscle loss: cachexia, sarcopenia, and inactivity. *Am J Clin Nutr* 2010;91:1123S-7S.
17. Kortebein P, Symons TB, Ferrando A, et al. Functional impact of 10 days of bed rest in healthy older adults. *J Gerontol A Biol Sci Med Sci* 2008;63:1076-81.
18. Covinsky KE, Palmer RM, Fortinsky RH, et al. Loss of independence in activities of daily living in older adults hospitalized with medical illnesses: increased vulnerability with age. *J Am Geriatr Soc* 2003;51:451-8.
19. Volpato S, Onder G, Cavalieri M, et al. Characteristics of non-disabled older patients developing new disability associated with medical illnesses and hospitalization. *J Gen Intern Med* 2007;22:668-74.
20. Schweickert WD, Pohlman MC, Pohlman AS, et al. Early physical and occupational therapy in mechanically ventilated, critically ill patients: a randomised controlled trial. *Lancet* 2009;373:1874-82.
21. Morris PE, Goad A, Thompson C, et al. Early intensive care unit mobility therapy in the treatment of acute respiratory failure. *Crit Care Med* 2008;36:2238-43.
22. de Morton NA, Keating JL, Jeffs K. The effect of exercise on outcomes for older acute medical inpatients compared with control or alternative treatments: a systematic review of randomized controlled trials. *Clin Rehabil* 2007;21:3-16.
23. Malone D. Bed rest, deconditioning, and hospital-acquired neuromuscular disorders. In: Malone D, Bishop Lindsay K, eds. *Physical therapy in acute care: a clinician's guide*, Thorofare: Slack Books; 2006:93-110.
24. Bayliss EA, Ellis JL, Powers JD, Gozansky W, Zeng C. Using self-reported data to segment older adult populations with complex care needs. *EGEMS (Wash DC)* 2019;7:12.
25. Jette DU, Stilphen M, Ranganathan VK, Passek SD, Frost FS, Jette AM. Validity of the AM-PAC "6-clicks" inpatient daily activity and basic mobility short forms. *Phys Ther* 2014;94:379-91.
26. Jette DU, Stilphen M, Ranganathan VK, Passek SD, Frost FS, Jette AM. AM-PAC "6-clicks" functional assessment scores predict acute care hospital discharge destination. *Phys Ther* 2014;94:1252-61.
27. Menendez ME, Schumacher CS, Ring D, Freiberg AA, Rubash HE, Kwon YM. Does "6-clicks" day 1 postoperative mobility score predict discharge disposition after total hip and knee arthroplasties? *J Arthroplasty* 2016;31:1916-20.
28. Covert S, Johnson JK, Stilphen M, Passek S, Thompson NR, Kazan I. Use of the activity measure for post-acute care "6 clicks" basic mobility inpatient short form and National Institutes of Health Stroke Scale to predict hospital discharge disposition after stroke. *Phys Ther* 2020;100:1423-33.
29. Bowles KH, Ratcliffe SJ, Holmes JH, et al. Using a decision support algorithm for referrals to post-acute care. *J Am Med Dir Assoc* 2019;20:408-13.
30. Bowles KH, Ratcliffe SJ, Naylor MD, Holmes JH, Keim SK, Flores EJ. Nurse generated EHR data supports post-acute care referral decision making: development and validation of a two-step algorithm. *AMIA Annu Symp Proc* 2017;2017:465-74.
31. Stucki G, Kostanjsek N, Ustun B, Ewert T, Cieza A. Applying the ICF to rehabilitation medicine editor. In: Frontera W, ed. *DeLisa's physical medicine and rehabilitation principles and practice*, Philadelphia: Lippencott Williams & Wilkins; 2010:301-24.
32. Haley SM, Coster WJ, Andres PL, et al. Activity outcome measurement for postacute care. *Med Care* 2004;42:149-61.
33. Jette AM, Haley SM, Tao W, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. *Phys Ther* 2007;87:385-98.
34. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *J Clin Epidemiol* 2006;59:1174-82.
35. Borson S, Scanlan JM, Chen P, Ganguli M. The Mini-Cog as a screen for dementia: validation in a population-based sample. *J Am Geriatr Soc* 2003;51:1451-4.
36. Cheville AL, Yost KJ, Larson DR, et al. Performance of an item response theory-based computer adaptive test in identifying functional decline. *Arch Phys Med Rehabil* 2012;93:1153-60.
37. Segall DO. Multidimensional adaptive testing. *Psychometrika* 1996;61:331-54.
38. Segall DO. General ability measurement: an application of multidimensional item response theory. *Psychometrika* 2001;66:79-97.
39. Wang C, Chang H, Boughton K. Deriving stopping rules for multidimensional computerized adaptive testing. *Appl Psychol Meas* 2013;37:99-122.
40. Wang WC, Chen PH, Cheng YY. Improving measurement precision of test batteries using multidimensional item response models. *Psychol Methods* 2004;9:116-36.
41. Reckase MD. The difficulty of test items that measure more than one ability. *Appl Psychol Meas* 1985;9:401-12.
42. Reckase MD, McKinley R. The discriminating power of items that measure more than one dimension. *Appl Psychol Meas* 1991;15:361-73.
43. Veldkamp BP, van der Linden WJ. Multidimensional adaptive testing with constraints on test content. *Psychometrika* 2002;67:575-88.
44. Wang C, Chang H. Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika* 2011;76:363-84.
45. Wang C, Chang H, Boughton K. Kullback-Leibler information and its application in multidimensional adaptive tests. *Psychometrika* 2011;76:13-39.
46. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res* 2007;16(Suppl 1):33-42.

47. Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multi-ethnic society. Overview to the special issue. *Med Care* 2006;44(11 Suppl 3):S3-4.
48. Teresi JA, Ramirez M, Lai JS, Silver S. 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. *Psychol Sci Q* 2008;50:538.
49. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. *Psychol Sci Q* 2009;51:148-80.
50. Wang Z, Weiss DJ, Wang C. DIF-CAT: doubly adaptive CAT using subgroup information to improve measurement precision. Niigata, Japan: International Association for Computerized Adaptive Testing; 2017.
51. Wang C, Fan Z, Chang H, Douglas J. A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J Educ Behav Stat* 2013;38:381-417.
52. Cheville AL, Wang C, Ni P, Jette AM, Basford JR. Age, sex, and symptom intensity influence test taking parameters on functional patient-reported outcomes. *Am J Phys Med Rehabil* 2014;93:931-7.
53. Wang C, Chang H, Douglas J. The linear transformation model with frailties for the analysis of item response times. *Br J Math Stat Psychol* 2013;66:148-68.
54. Fan Z, Wang C, Chang H, Douglas J. Utilizing response time distributions for item selection in CAT. *J Educ Behav Stat* 2012;37:579-600.
55. Kim-Kang G, Weiss DJ. Adaptive measurement of individual change. *Z Psychol* 2008;216:49-58.
56. Cronbach LJ, Furby L. How we should measure "change" – or should we? *Psychol Bull* 1970;74:68-80.
57. Quatrano LA, Cruz TH. Future of outcomes measurement: impact on research in medical rehabilitation and neurologic populations. *Arch Phys Med Rehabil* 2011;92(10 Suppl):S7-11.
58. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity –establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health* 2011;14:978-88.
59. Serlin RC, Mendoza TR, Nakamura Y, Edwards KR, Cleeland CS. When is cancer pain mild, moderate or severe? Grading pain severity by its interference with function. *Pain* 1995;61:277-84.
60. Paul SM, Zelman DC, Smith M, Miaskowski C. Categorizing the severity of cancer pain: further exploration of the establishment of cutpoints. *Pain* 2005;113:37-44.
61. Li KK, Harris K, Hadi S, Chow E. What should be the optimal cut points for mild, moderate, and severe pain? *J Palliat Med* 2007;10:1338-46.
62. Bergstrom N, Braden B, Kemp M, Champagne M, Ruby E. Predicting pressure ulcer risk: a multisite study of the predictive validity of the Braden Scale. *Nurs Res* 1998;47:261-9.
63. O'Dell M, Barr K, Spanier D, Warnick RE. Functional outcome of inpatient rehabilitation in persons with brain tumors. *Arch Phys Med Rehabil* 1998;79:1530-4.
64. Huang ME, Cifu DX, Keyser-Marcus L. Functional outcomes in patients with brain tumor after inpatient rehabilitation: comparison with traumatic brain injury. *Am J Phys Med Rehabil* 2000;79:327-35.
65. Granger CV, Cotter AC, Hamilton BB, Fiedler RC. Functional assessment scales: a study of persons after stroke. *Arch Phys Med Rehabil* 1993;74:133-8.
66. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: a study of persons with multiple sclerosis. *Arch Phys Med Rehabil* 1990;71:870-5.
67. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;19:716-23.
68. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461-4.
69. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 1981;46:443-59.
70. Templin TN, Hays RD, Gershon RC, et al. Introduction to patient-reported outcome item banks: issues in minority aging research. *Expert Rev Pharmacoecon Outcomes Res* 2013;13:183-6.
71. Roussos L, Stout W. A multidimensionality-based DIF analysis paradigm. *Appl Psychol Meas* 1996: 20.
72. Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006;44(11 Suppl 3):S182-8.
73. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 1993;58:525-43.
74. Meredith W, Teresi JA. An essay on measurement and factorial invariance. *Med Care* 2006;44(11 Suppl 3):S69-77.
75. Reise SP, Moore T, Maydeu-Olivares A. Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educ Psychol Meas* 2011;71:684-711.
76. Reise SP. The rediscovery of bifactor measurement models. *Multivariate Behav Res* 2012;47:667-96.
77. Cai L, Thissen D, du Toit SHC. IRTPRO: flexible, multidimensional, multiple categorical IRT modeling. Chicago: Scientific Software International Inc; 2011.
78. Raju N, van der Linden W, Fleer P. IRT-based internal measures of differential functioning of items and tests. *Appl Psychol Meas* 1995;19:353-68.
79. Wainer H, Sireci S, Thissen D. Differential testlet functioning: definitions and detection. *J Educ Meas* 1991;28:197-219.
80. Chang H, Mazzeo J. The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika* 1994;39:391-404.
81. Collins W, Raju N, Edwards J. Assessing differential item functioning in a satisfaction scale. *J Appl Psychol* 2000;85:451-61.
82. Morales L, Flowers C, Gutiérrez P, Kleinman M, Teresi JA. Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. *Med Care* 2006;44(11 Suppl 3):S143-51.
83. Orlando-Edelen M, Thissen D, Teresi J, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: applications to the Mini-Mental State Examination. *Med Care* 2006;44(11 Suppl 3):S134-542.
84. Steinberg L, Thissen D. Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychol Methods* 2006;11:402-15.
85. Teresi J, Ocepek-Welikson K, Kleinman M, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measure of physical functioning ability and general distress. *Qual Life Res* 2007;16:43-68.
86. Lord F, Novick M. Statistical theories of mental test scores. Reading: Addison-Wesley Publishing Co; 1968.
87. Oshima T, Kushubar S, Scott J, Raju N. DFIT8 for Windows user's manual: differential functioning of items and tests. St Paul: Assessment Systems Corporation; 2008.
88. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front Psychol* 2016;7:109.
89. DeMars C. Modeling DIF for simulations: continuous or categorical secondary trait? *Psychol Test Assess Model* 2015;57:279-300.

90. van der Linden WJ. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 2007;72:287.
91. Tao W, Haley SM, Coster WJ, Ni P, Jette AM. An exploratory analysis of functional staging using an item response theory approach. *Arch Phys Med Rehabil* 2008;89:1046-53.
92. Hays RD, Spritzer KL, Thompson WW, Cella DUS. general population estimate for "excellent" to "poor" self-rated health item. *J Gen Intern Med* 2015;30:1511-6.
93. Bergstrom N, Braden BJ, Laguzza A, Holman V. The Braden Scale for predicting pressure sore risk. *Nurs Res* 1987;36:205-10.
94. Wang N. Use of the Rasch IRT model in standard setting: an item-mapping method. *J Educ Meas* 2003;40:231-53.
95. Reckase M. A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educ Meas* 2006;25:4-18.
96. MacCann R, Stanley G. The use of Rasch modeling to improve standard setting. *Pract Asses Res Eval* 2006;11.
97. Wang C, Weiss D, Shang Z. Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika* 2019;84:749-71.
98. Jette DU, Stilphen M, Ranganathan VK, Passek S, Frost FS, Jette AM. Interrater reliability of AM-PAC "6-clicks" basic mobility and daily activity short forms. *Phys Ther* 2015;95:758-66.
99. Hoyer EH, Young DL, Klein LM, et al. Toward a common language for measuring patient mobility in the hospital: reliability and construct validity of interprofessional mobility measures. *Phys Ther* 2018;98:133-42.
100. Hoyer EH, Friedman M, Lavezza A, et al. Promoting mobility and reducing length of stay in hospitalized general medicine patients: a quality-improvement project. *J Hosp Med* 2016;11:341-7.
101. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol* 2008;61:17-33.
102. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware Jr. JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol* 2014;67:516-26.
103. Bruce B, Fries JF, Ambrosini D, et al. Better assessment of physical function: item improvement is neglected but essential. *Arthritis Res Ther* 2009;11:R191.