JRRD Volume 45, Number 9, 2008
Pages 1385–1400
Journal of Rehabilitation Research & Development

Rasch analysis of Minimum Data Set mandated in skilled nursing facilities

Ying-Chih Wang, PhD;[1*] Katherine L. Byers, PhD;[2] Craig A. Velozo, PhD, OTR/L[3]
[1]Sensory Motor Performance Program, Rehabilitation Institute of Chicago, Chicago, IL; [2]Department of Clinical Administration and Rehabilitation Counseling, Texas Tech University Health Science Center, Lubbock, TX; [3]Department of Veterans Affairs Health Services Research and Development and Rehabilitation Research and Development Services, Rehabilitation Outcomes Research Center, North Florida/South Georgia Veterans Health System, and Department of Occupational Therapy, University of Florida, Gainesville, FL

Abstract—This study examined the construct validity of the physical functioning and cognition scales of the federally mandated Minimum Data Set (MDS) via Rasch analysis. We performed a secondary analysis of retrospective MDS data collected by the Department of Veteran Affairs Austin Automation Center. Items demonstrated unidimensionality and represented two separate constructs: physical functioning and cognition. The physical functioning items showed good psychometric properties and covered a wide range of residents' physical functioning, with the spread of items efficiently discriminating residents' performance into different physical functioning strata. In contrast, the psychometric properties of the cognition items were less sound, had severe ceiling effects, and did not efficiently discriminate residents' performance into different cognition strata. The constructed validity of the physical functioning and cognition scales of the MDS, in general, was supported. Future investigations into the function of the rating scale and the necessity of adding challenging items are recommended.

Key words: activities of daily living, assessment, cognition, construct validity, MDS, patient outcomes, physical functioning, Rasch analysis, rehabilitation, skilled nursing facilities.

INTRODUCTION

According to the U.S. Census Bureau, more than 36.6 million individuals in the United States are age 65 and over and this population is projected to triple by 2050 [1]. The increasing age of the population is accompanied by a greater number of people living with chronic disease, including functional limitations and disabilities [2–3]. Based on the U.S. Centers for Disease Control and Prevention's health statistics report, 34 percent of the elderly population have activity limitations caused by chronic conditions and 6.3 percent need help with personal care [4]. Cognitive impairment, which is also common among elderly people, was associated with a higher risk of functional decline and poor functional recovery [5]. Cognitive impairment also has an effect on the ability to perform

Abbreviations: ADL = activity of daily living, CMS = Centers for Medicare & Medicaid Services, DIF = differential item functioning, FA = factor analysis, FIM = Functional Independence Measure™, MDS = Minimum Data Set, MNSQ = mean-square, OASIS = Outcome and Assessment Set, RAI = Resident Assessment Instrument, RAP = Resident Assessment Protocol, SNF = skilled nursing facility, VA = Department of Veterans Affairs.

[*]Address all correspondence to Ying-Chih Wang, PhD; Sensory Motor Performance Program, Rehabilitation Institute of Chicago, 345 East Superior, Room 1312, Chicago, IL 60611; 312-238-4109; fax: 312-238-2208.
Email: inga-wang@northwestern.edu
DOI:10.1682/JRRD.2007.11.0184

activities of daily living (ADLs) and is associated with an increased cost of care in the elderly population [6].

Nursing homes are a critical environment for tracking the healthcare status of the elderly population. Individuals who cannot take care of themselves because of physical, emotional, or mental problems may choose or be placed in skilled nursing facilities (SNFs). Currently, 1.6 million residents live in nursing homes and their average length of stay is approximately 892 days [7]. More than 90 percent of current residents are 65 years of age or older, and most residents require assistance in multiple ADLs [8]. Reports estimate that about 40 percent of nursing home residents need help with eating and more than 90 percent require assistance with bathing [9].

To improve the quality of care in SNFs, the Centers for Medicare & Medicaid Services (CMS) developed a Resident Assessment Instrument (RAI) in 1990 to assess and plan care for residents in SNFs [10]. As of 1998, with regulations and the introduction of a prospective payment system, SNFs are required to complete and transmit RAI data to the state for all residents.

As a central assessment core in the RAI, the Minimum Data Set (MDS) covers 18 clinically important domains [11]. With approximately 450 items in a fully comprehensive assessment (about half of which need to be completed during quarterly assessments), the MDS gathers abundant resident background information for designing care plans, evaluating quality of care, and further monitoring the impact of policy changes [10,12].

Numerous studies have investigated the psychometric properties of the MDS. Several reliability and validity properties of the MDS, including interrater reliability [12–13], test-retest reliability [14], rater agreement [12], concurrent validity [15–16], responsiveness [16], and dimensionality [12], have been reported in the literature. Many studies support the reliability and clinical utility of the MDS items and suggest MDS data should be used for research purposes [12–13,17–19]. Hawes et al. reported that MDS items met a standard for excellent reliability in areas of functional status such as ADLs, continence, cognition, and diagnoses, with intraclass correlation coefficients of 0.7 or higher [20]. Casten and colleagues found high correlations between the raters for each index (e.g., cognition: $r = 0.80$, ADLs: $r = 0.99$) [12]. However, the MDS has been criticized for its assessment procedures and lack of reliability in rating [12]. The care coordinator who oversees the completion of the MDS may either ask questions of other staff orally or have other health professionals actually rate relevant items [12,18,20].

Through the Nursing Home Quality Initiative, which was initiated in November 2002, the CMS continues to work with measurement experts to improve the quality of measures for SNFs. While not all quality indicators require data from multiple items or use a rating scale, assessments based on combining items with rating scales can be useful in monitoring patient outcomes. For a better understanding of healthcare instruments, researchers must document the psychometric properties of these assessments. Findings from these analyses may suggest that a revision of the instrument is necessary, which is consistent with the commitment from the CMS to continue to revise and improve the RAI for care planning. These psychometric analyses all focus on reliability and validity at the total score level of the MDS. An alternative approach is to inspect the rating scale structure and examine the underlying psychometrics of the MDS at the item level.

A myriad of studies have used Rasch analysis to examine and refine instruments in the health-related field [21–24]. However, no published studies have applied Rasch analysis to explore the psychometric properties of MDS items. As a step to build on the existing psychometric studies related to the MDS instrument, we conducted this study to assess the physical functioning and cognition domains of the MDS using Rasch analysis.

METHODS

Sample

This secondary data analysis was performed on data collected by the Department of Veterans Affairs (VA) Austin Automation Center from June 2002 to May 2003. Data were downloaded from the RAI-MDS database. The RAI-MDS database contains a core set of clinical and functional status elements (i.e., MDS), triggers, and 18 Resident Assessment Protocols (RAPs). The long-term care programs may be provided by a VA medical center or a non-VA facility that may be financed by VA, Medicare, Medicaid, private insurance, and/or out-of-pocket. State veterans' homes, which are funded by the VA and also participate in Medicare and Medicaid, are required to collect residents' information for care planning and transmit MDS data to the CMS.

Inclusion criteria were patients who (1) had a stroke, amputation, or orthopedic impairment code and (2) had

no missing values in any of the MDS physical functioning or cognition items. The final data set comprised a total sample of 654 veterans' records, with 302 stroke, 113 amputation, and 239 orthopedic impairment patients. The average age of this sample was 68.2 ± 12.5 years, 96.6 percent were male, 74.2 percent were white, and 46.7 percent were married. All patients were from VA hospitals, except two who were from VA nursing homes. **Table 1** provides the subject baseline demographic characteristics and information on impairment categories.

Table 1.
Demographic characteristics (n = 654).

Characteristic	*n* (%)
Sex*	
Male	630 (96.6)
Female	22 (3.4)
Race/Ethnicity	
White	485 (74.2)
Black	120 (18.3)
Hispanic	26 (4.0)
Native American	8 (1.2)
Asian	2 (0.3)
Other	5 (0.8)
Missing	8 (1.2)
Impairment Group	
Stroke	
Left Body Involvement	140 (21.4)
Right Body Involvement	134 (20.5)
Bilateral Involvement	7 (1.1)
No Paresis	8 (1.2)
Other Stroke	13 (2.0)
Amputation (Lower Limb)	
Unilateral Above Knee	30 (4.6)
Unilateral Below Knee	71 (10.9)
Bilateral Above Knee	1 (0.2)
Bilateral Above/Below Knee	2 (0.3)
Bilateral Below Knee	9 (1.4)
Orthopedic	
Unilateral Hip Fracture	31 (4.7)
Bilateral Hip Fractures	1 (0.2)
Femur Fracture	5 (0.8)
Pelvic Fracture	3 (0.5)
Major Multiple Fractures	6 (0.9)
Unilateral Hip Replacement	74 (11.3)
Unilateral Knee Replacement	84 (12.8)
Bilateral Knee Replacements	2 (0.3)
Other Orthopedic	33 (5.0)

*Sex variable has 2 missing values.

This study was approved by the institutional review board at the University of Florida and the VA Subcommittee on Human Studies. Access to VA MDS data was approved by the VA Veterans Health Administration.

Resident Assessment Instrument-Minimum Data Set

The physical functioning items are embedded in section G Physical Functioning and Structural Problems of the MDS. Items are intended to measure residents' ADLs, such as bed mobility, transferring, dressing, locomotion, eating, hygiene, and bathing. All items have a 5-point rating scale ranging from 0 (independent) to 4 (total dependence), with lower scores representing higher levels of performance. If the activity did not occur during the entire 7 days, the rater is instructed to score an 8 (activity did not occur). In this study, instead of treating the MDS rating scale score 8 as missing, we recoded it to 4 (total dependence). This recoding was based on the rationale that the most likely explanation of a basic ADL not being observed during the entire observation period is inability to perform the task [25–26].

The cognition items are embedded in section B (Cognitive Patterns) and section C (Communication/Hearing Patterns) of the MDS. These items are used for evaluating residents' memory, perception/awareness, cognitive skills for daily decision making, and communication performance. Unlike the physical functioning scale, in which all items share the same 5-point rating scale structure, each cognition item has its own rating scale category and definition. Two memory and four recall items are dichotomous items (0–1); seven items have a 3-point rating scale (0–2); and three items have a 4-point rating scale (0–3).

Administration of Minimum Data Set

When the resident is admitted to a facility, the Registered Nurse Assessment Coordinator and the interdisciplinary team (e.g., physician, nursing assistant, social worker, and therapist) have a 14-day observation period in which to complete the admission assessment. After the initial MDS assessment is completed, the RAP is reviewed to identify the resident's strengths, problems, and needs for a future care plan. After the initial comprehensive assessment, a quarterly assessment is mandated 90 days later, and an annual comprehensive assessment is required to be completed no more than 366 days from the date of the prior comprehensive assessment. Furthermore, the staff must complete additional assessments when a resident is discharged or has significant change. Because of the large

amount of paperwork, some facilities hire MDS contract nurses to complete the records based on information provided in the residents' charts [16].

Rasch Analysis

We performed Rasch analysis using the Winsteps® program (version 3.16, Winsteps; Chicago, Illinois) [27] to evaluate the MDS physical functioning and cognition items. The Rasch model (also called a one-parameter logistic item response theory model) is a probabilistic, mathematical model that assumes the probability of passing an item depending on the relationship between a person's ability and an item's difficulty. It is based on the concept that data must conform to some reasonable hierarchy of less than/more than on a single continuum of interest [28]. By inspecting persons' responses to items that are relatively harder or easier to endorse, the Rasch model establishes the item difficulty hierarchical order from the easiest to the most challenging tasks according to a specific sample.

The Rasch model has several advantages over traditional classical test theory. First, besides exploring the data at the test level (e.g., reliability and validity), the Rasch model can inspect the data at the item level, including item difficulty, rating scale structure, and whether response patterns fit the expected measurement pattern. Second, the item parameters are invariant regardless of whether a sample subgroup with higher or lower ability is used to estimate the item parameter (sample free). Regardless of whether the test was given to a group of examinees with low, high, or a variety of ability levels, the item parameter estimation remains the same. Third, the person ability is estimated independently of the particular set of items that are administered to the examinee (scale free). Regardless of whether a set of harder or easier items is given to the examinees, the examinees' ability estimations remain the same. Lastly, items in the instrument are reported on the same scale as person-ability scores, which allows for investigation of the extent to which the item difficulties match the person abilities of a particular sample.

Because of the challenge of using a different rating scale for the cognition items and in the belief that the rating scale in the physical functioning subscales shared different structures as well (i.e., not sharing equal distance between individual rating scale categories), we used a partial credit model [29] to analyze the data. In addition, for easier interpretation of the results, the MDS codes were reversed prior to performing the Rasch analysis so that a higher rating indicated higher function.

Analytical Procedure

Dimensionality

Many measurement experts believe that meaningful "objective" measurement can only be achieved if each item contributes to the measurement of a single attribute [30]. Therefore, factor analysis (FA) was used to examine the dimensionality of the instrument. The unidimensionality of the scale is determined by interpreting the factor loading matrix (the correlations between the original variables and the common factors) and the percent of variance explained by each factor. After initial FA without rotation, we used FA varimax (orthogonal transformation) and FA promax (oblique transformation) as follow-up analyses for better interpretability of the results.

How Well Data Fit Model

Fit statistics were performed to investigate whether the response patterns on the physical functioning and cognition scales fit the Rasch measurement model. A fit statistic index calculates the ratio of the observed variance divided by the expected variance, with an expected value of 1 and a range from 0 to positive infinity. A mean-square (MNSQ) fit value of $1 + X$ indicates the observed variance contains $100 \times X$ percent more variation than the model predicted [28]. Based on Linacre [31] and Smith et al. [32], we used a reasonable item MNSQ fit statistic of 1.2 given that we had a sample size of approximately 600. A MNSQ fit statistic higher than 1.2 indicates that the item response pattern has more variance than the model expected. There are two kinds of fit statistics: the infit statistic is a weighted index and is more sensitive to the response pattern of items targeted to the person's estimated ability level, and the outfit statistic is an unweighted index that computes the overall misfit of personal responses [33].

Item Difficulty Hierarchy

The empirical item difficulty hierarchical order produced by the Rasch analysis can be used as evidence of construct validity to the theoretical base of the instrument. Item difficulty hierarchical order was inspected via the estimated item difficulty calibrations, which are expressed in logits, with higher positive values indicating a more challenging task.

Person Item Match-Targeting

In Rasch analysis, both person ability and item difficulty are expressed on a common metric. The extent to which the items are of appropriate difficulty for the sample can be examined by comparing the sample ability distribution to the item difficulty distribution. Ceiling effects can be depicted by a lack of items for persons of high ability, and floor effects can be depicted by a lack of items matching persons of low ability. Furthermore, clusters of items or gaps between items (no items within a range of a person-ability level) may indicate a redundancy of items or the need to add items within an instrument.

Separation Index

The precision of measurement depends on how well the items of an instrument separate individuals of different ability levels. The person-separation index is an estimate of how well the instrument can differentiate persons on the measured variable. A separation index above 2 is required to attain the desired reliability level of at least 0.8 [34]. The person-separation index (G) can be further computed into the number of statistically distinct person strata identified by the formula $[(4G + 1) / 3]$ [35]. This value indicates how many distinct levels of person ability can be statistically differentiated in ability strata with centers three measurement errors away [36].

Rating Scale Structure

The rating scale structure was initially examined by inspecting the frequency count for each response option. Categories with low frequencies indicate that the performance level/rating scale can be assigned to the respondent only in rare situations or with a narrowly defined scope. We further used Linacre's rating scale criteria to exam the rating scale structure [37].

Differential Item Functioning

Differential item functioning (DIF) analysis can be used to examine whether items function similarly across different groups and identify items that appear to be too easy or difficult after controlling for the ability levels of the compared groups. In this study, we used the DIF method [38] to explore whether items on the MDS performed similarly across three different diagnostic groups (individuals with stroke, amputation, and orthopedic impairment) and among patients with right versus left hemiparesis. Since DIF analysis is a series of pairwise *t*-tests, which equal DIF contrast divided by the joint standard error of the two DIF measures, a critical value (*p*-value) of 0.01 was used to determine the statistical significance of the DIF.

RESULTS

Dimensionality

Initially, FA without rotation was performed. Within a conjoint run of all functional items (physical functioning and cognition items), the results of FA showed that five factors had eigenvalues greater than 1. The first five factors had eigenvalues equal to 11.9, 3.7, 1.9, 1.3, and 1.0, respectively, which explained approximately 41, 13, 7, 4, and 4 percent of the total variance, respectively. Results from the factor pattern revealed that for the first component, all items had positive loadings ranging from 0.41 to 0.80, which indicated a general construct measuring functional status. For the second factor, all cognition items had positive loadings (0.11–0.47) and all physical functioning items had negative loadings (–0.05 to –0.51), indicating two separate subconstructs were underlying the overall functional status domain. The third factor had relatively high factor loadings (>0.35) on six items that were indicators of delirium (i.e., easily distracted, periods of altered perception, restlessness, lethargy, disorganized speech, and mental function varies over the course of the day). Lastly, while three communication items showed high factor loadings on the fourth factor (0.37–0.52), two walking items demonstrated relatively high factor loadings on the fifth factor (0.53–0.54).

We then performed orthogonal transformation, followed by oblique transformation, in which factors are allowed to be correlated with each other. Factor patterns obtained from orthogonal and oblique transformation showed results similar to those just mentioned.

Rasch Analysis-Physical Functioning Items

Overall, the physical functioning items showed good psychometric properties. Person reliability (analogous to Cronbach alpha) was 0.89. With the criteria of 1.2 for the MNSQ fit statistics, two items (locomotion off unit and bladder) showed high infit statistics and five items (walking in corridor, walking in room, locomotion off unit, bowel, and bladder) demonstrated high outfit statistics (**Table 2**).

The physical functioning item difficulty calibrations are also presented in **Table 2**. The mean value of item difficulty calibrations ranged from –1.37 to 1.49 logits,

Table 2.
Physical functioning item statistics listed by item difficulty order from most to least difficult.

Physical Functioning Item	Measure*	Error	Infit MNSQ	Outfit MNSQ	Score CORR	Average Measure for Each Rating Scale				
						4	3	2	1	0
Walking in Corridor	1.49	0.04	1.28	2.16	0.68	–0.17	0.16	0.89	1.43	2.02
Walking in Room	1.22	0.04	1.28	2.01	0.69	–0.26	–0.24	0.66	1.23	1.96
Bathing	1.11	0.05	0.91	0.85	0.77	–1.25	0.46	1.11	1.59	2.05
Locomotion off Unit	0.61	0.04	1.51	1.77	0.67	–0.61	–0.14	0.56	1.29	1.44
Dressing	0.14	0.05	0.82	0.84	0.79	–1.75	–0.54	0.54	1.15	1.78
Toileting	0.12	0.05	0.56	0.53	0.82	–1.70	–0.51	0.40	1.11	1.78
Transfer	–0.07	0.05	0.72	0.69	0.80	–1.80	–0.59	0.24	0.93	1.73
Hygiene	–0.27	0.05	0.78	0.78	0.78	–2.08	–0.73	0.19	0.96	1.51
Locomotion on Unit	–0.35	0.05	1.24	1.23	0.69	–1.35	–0.94	–0.03	0.87	1.15
Bowel	–0.78	0.05	0.98	1.70	0.68	–1.68	–0.84	–0.49	0.12	0.97
Bladder	–0.82	0.05	1.63	2.29	0.62	–1.60	–0.53	–0.27	0.30	0.90
Bed Mobility	–1.03	0.05	1.12	1.04	0.70	–2.58	–1.05	–0.16	0.45	1.12
Eating	–1.37	0.06	0.99	1.05	0.70	–2.22	–1.62	–0.69	0.08	1.09
Mean ± SD	0.00 ± 0.87	0.05 ± 0.00	1.06 ± 0.30	1.30 ± 0.58	0.72 ± 0.06	–1.47 ± 0.73	–0.55 ± 0.54	0.23 ± 0.54	0.89 ± 0.50	1.50 ± 0.42

*Item difficulty calibration.
CORR = item-total correlation, MNSQ = mean-square, SD = standard deviation.

with an average of 0.05 logits error associated with parameter estimations. The range between the measure for the lowest MDS rating score of the easiest item and the measure for the highest MDS rating score for the hardest item was –2.22 to 2.02. Two walking items (walking in corridor and walking in room) and the bathing item were the most challenging items. Alternatively, eating, bed mobility, bladder, and bowel were the easiest items. Items such as toileting, dressing, transferring, and hygiene represented average difficulty items. Additionally, the score correlations (point-biserial correlations) between the individual item responses and the total test score were moderate to high (r = 0.62–0.82).

Figure 1 illustrates the relationship of the sample measure distribution (left) with the hierarchical order of the physical functioning items (right). Linear measures, in logits, are represented along the central axis. The distribution of person-ability estimations (higher values representing higher ability and lower values representing lower ability) were normally distributed with a slight ceiling effect. About 6.1 percent of the sample received a maximum measure. The mean sample ability level (M on the left) of 0.58 ± 1.76 logits matched well with the mean item difficulty of the MDS items (M on the right) of 0.00 ± 0.87 logits. With a person-separation index (G) equal to 2.89, these physical functioning items separated person ability into 4.19 statistically distinct strata.

The rating scale structure was initially examined by inspecting the frequency count for each response category. **Figure 2** shows the frequency count of responses from 0 (independent) to 4 (total dependence) and the additional rating scale response of 8 (activity did not occur during the entire 7 days). Three rating scale categories are presented on each graph to simplify the presentation. On the x-axis, 13 physical functioning items were listed and ordered from the easiest (eating) to the most challenging item (walking in corridor) (from left to right). The y-axis is the frequency count of the rating scale response. As items increased in difficulty, the frequency counts of 0 (independent) decreased as expected. In general, the frequency count of 1 (supervision) maintained a relatively low frequency count independent of the difficulty of the item. Items at the average difficulty level had a relatively high frequency count, being scored with 2 (limited assistance) or 3 (extensive assistance). A relatively high frequency count was noted for limited assistance with the dressing item, and a particularly high frequency count was found for extensive assistance with the bathing item. The trend of the frequency count of 4 (total dependence) did not monotonically increase as the difficulty of the task increased. The frequency count of total dependence was much higher for bathing and locomotion off unit (as expected for more difficult items), yet was very low for the two most challenging

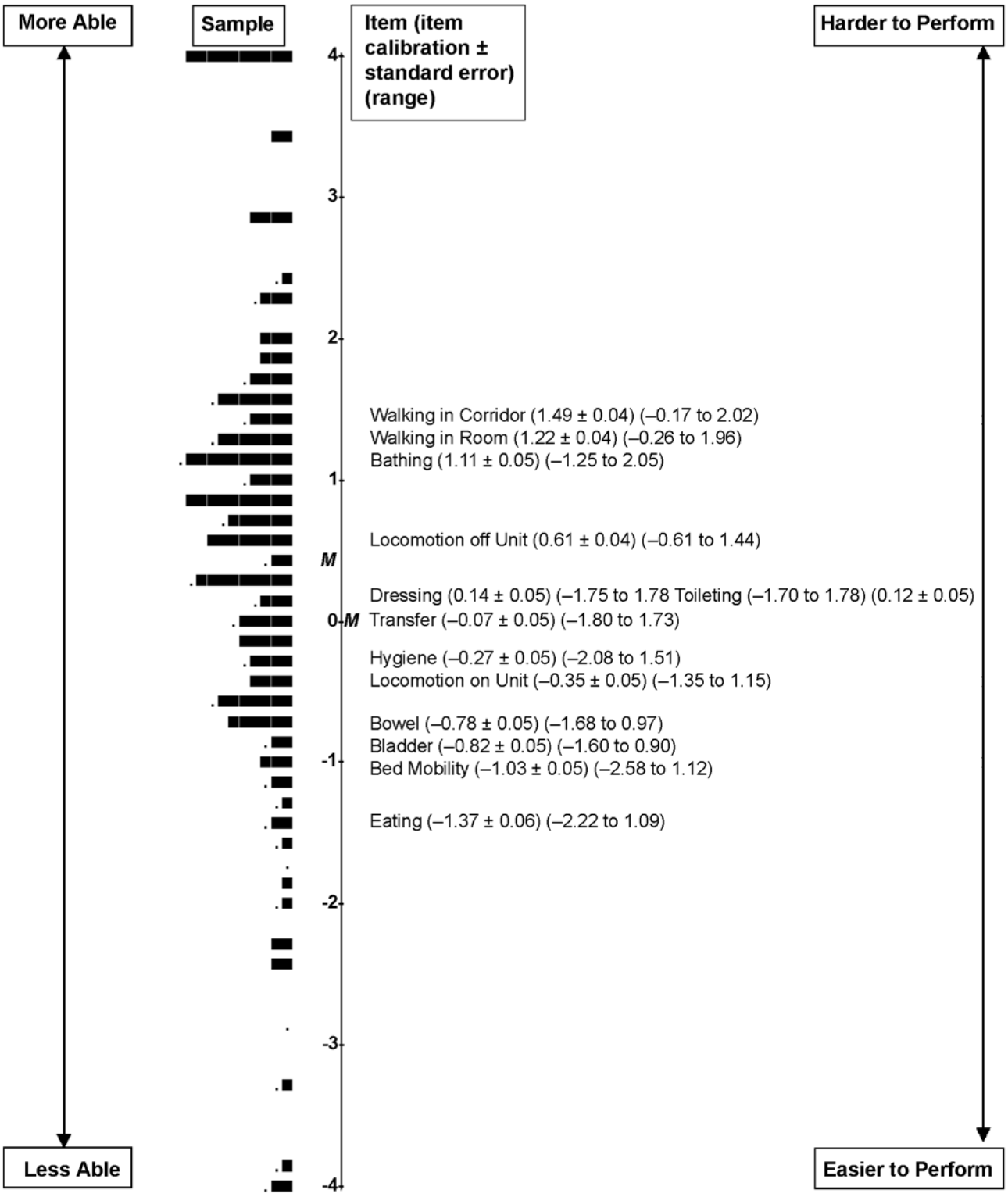

Figure 1.
Physical functioning subscale of Minimum Data Set: Person score distribution (left) and item difficulty hierarchy map (right). Each "■" indicates 4 persons, and each "." indicates 1 person. *M* represents the mean of person ability measures (left) and item difficulty calibrations (right).

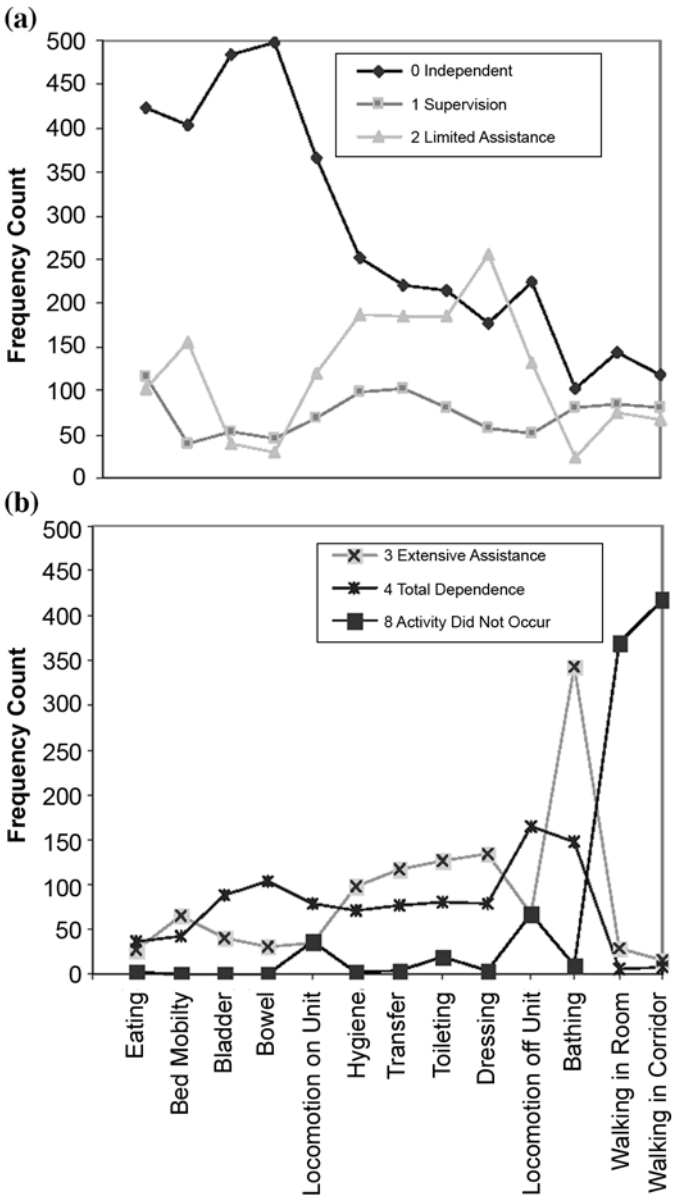

Figure 2.
Rating scale frequency count of physical functioning items. **(a)** Rating scale categories from 0 to 2. **(b)** Rating scale categories 3, 4, and 8. Items have been ordered by difficulty from least to most difficult along *x*-axis.

items, walking in room and walking in corridor. For the special rating scale response of 8 (activity did not occur during the entire 7 days), the frequency count was low across all items, except the two walking items: walking in room and walking in corridor.

Most of the physical functioning items met Linacre's essential criteria for optimizing rating scale categories [37]. All the rating scale categories had at least 10 observations. The average measures for each rating scale structure advanced monotonically within the category. Four items (eating, bladder, bowel, walking in corridor) had one rating scale category that showed misfit (outfit MNSQ greater than the criterion of 2). The locomotion off unit item had two rating scale categories that showed misfit.

After performing the DIF analysis across the three impairment groups, we found several items to have significant DIF. In comparing patients with stroke to those with amputation, we found that patients with stroke had more difficulty in hygiene, eating, and bathing ($p < 0.001$) and bladder ($p = 0.002$) and that patients with amputation had more difficulty walking (walking in room and walking in corridor) ($p < 0.001$). In comparing patients with stroke to those with orthopedic impairment, we found that patients with stroke had more difficulty in hygiene, eating, and continence (bowel) ($p < 0.001$). On the other hand, patients with orthopedic impairment had more difficulty walking in room, walking in corridor, transferring, and bed mobility ($p < 0.001$). Lastly, when comparing patients with amputation to those with orthopedic impairment, we found that patients with amputation experienced more challenges in walking (walking in room and walking in corridor) ($p < 0.001$) and that patients with orthopedic impairment had more difficulty with bed mobility and dressing ($p < 0.001$). No significant DIF was found when patients with stroke with left hemiparesis versus stroke with right hemiparesis were compared.

Rasch Analysis-Cognition Items

The psychometric characteristics of the MDS cognition items were slightly less sound than the psychometric characteristic of the physical functioning items. Person reliability (analogous to Cronbach alpha) was 0.68. Three cognition items (making self understood, speech clarity, and disorganized speech) showed infit statistics that just exceeded the critical value of 1.2, although three items, recall in nursing home, speech clarity, and lethargy, showed significantly high outfit statistics.

Table 3 presents the item difficulty estimates, infit/outfit statistics, and score correlations of the cognition items. The mean value of item difficulty calibrations ranged from –1.71 to 2.20 logits, with an average of 0.14 logits error associated with parameter estimations. The range between the measure for the lowest MDS rating score from the easiest item and the measure for the highest rating score from the hardest item was 0.42 to

Table 3.
Cognition item statistics listed by item difficulty order from most to least difficult.

Cognition Item	Measure*	Error	Infit MNSQ	Outfit MNSQ	Score CORR	Average Measure for Each Rating Scale			
						3	2	1	0
Short-Term Memory	2.20	0.13	0.88	0.81	0.75	—	—	1.31	4.27
Recall Staff Names/Faces	1.54	0.13	1.18	1.23	0.61	—	—	1.41	4.03
Recall Location of Own Room	1.24	0.14	0.88	0.83	0.67	—	—	0.99	4.03
Daily Decision Making	1.23	0.08	0.74	0.74	0.85	0.06	0.96	2.39	4.47
Recall Current Season	1.15	0.14	0.81	0.71	0.69	—	—	0.86	4.03
Long-Term Memory	0.96	0.14	0.77	0.69	0.68	—	—	0.76	3.99
Recall in Nursing Home	0.86	0.14	1.09	1.58	0.54	—	—	1.25	3.88
Making Self Understood	–0.11	0.09	1.26	1.09	0.64	–0.28	0.77	1.46	4.01
Speech Clarity	–0.33	0.12	1.31	1.78	0.47	—	0.98	1.43	3.82
Ability to Understand Others	–0.41	0.10	0.98	0.99	0.64	–0.31	0.35	1.23	3.98
Mental Function Varies	–1.02	0.15	0.97	0.68	0.49	—	0.13	0.68	3.75
Altered Perception/Awareness	–1.19	0.16	1.18	0.83	0.43	—	0.23	0.75	3.70
Restlessness	–1.29	0.16	1.05	0.97	0.43	—	0.57	0.55	3.70
Easily Distracted	–1.43	0.17	0.98	0.73	0.43	—	0.10	0.55	3.69
Disorganized Speech	–1.68	0.18	1.25	0.96	0.36	—	0.44	0.47	3.63
Lethargy	–1.71	0.19	1.11	1.74	0.32	—	0.42	1.09	3.62
Mean ± SD	0.00 ± 1.26	0.14 ± 0.03	1.03 ± 0.17	1.02 ± 0.36	0.56 ± 0.15	–0.18 ± 0.21	0.50 ± 0.32	1.07 ± 0.49	3.91 ± 0.24

*Item difficulty calibration.
CORR = item-total correlation, MNSQ = mean-square, SD = standard deviation.

4.27. Short-term memory, ability to recall, and daily decision making items formed the most challenging items along this construct. Communication items (making self understood, speech clarity, and ability to understand others) were the next most difficult items. Altered perception/awareness, easily distracted, disorganized speech, and lethargy were the easiest items. The score correlations (point-biserial correlations) between the individual item responses and the total test score ranged from 0.32–0.75, with score correlations low for those items associated with periodic disordered thinking/awareness and communication items (except the making self understood item) being fairly low (r = 0.32–0.49).

Figure 3 shows a map of the person-cognitive measures to the left and MDS item measures to the right. In contrast to the physical functioning items, which showed a good match between person measures and item measures, the cognition measure was "easy" for this sample. The average item difficulty (0.00 ± 1.26 logits, M to the right) was much lower than the average ability of the sample (3.49 ± 1.74 logits). Excluding persons with extreme data who obtained the total maximum score, the average ability of the sample was about 2.13 ± 1.35 logits (M to the left in the figure). The score distribution of the cognition subscale was highly skewed, with 47.5 percent of the sample showing perfect scores. With the person-separation index equal to 1.46, the cognition items distinguished persons into 2.28 statistically distinct strata.

Figure 4 shows the frequency count for each response category for the cognition items. On the x-axis, items were listed and ordered from the easiest (lethargy) to the most challenging (short-term memory) (from left to right). The y-axis is the frequency count of the rating scale category. The ceiling effect of the cognition subconstruct was evident in the rating scale frequency count, in which the majority of residents (68%–96%) were rated able or behavior not present (rating of 0) for their cognitive status on multiple cognition items. Meanwhile, several items (8 out of 16 items) had very low frequency counts (<10) in the rating scale category, indicating that subjects had severely impaired cognition status for those items.

Most of the cognition items met Linacre's essential criteria for optimizing rating scale categories [37]. The average measures for each rating scale structure advanced monotonically. Three items (speech clarity, restlessness, disorganized speech) had one rating scale

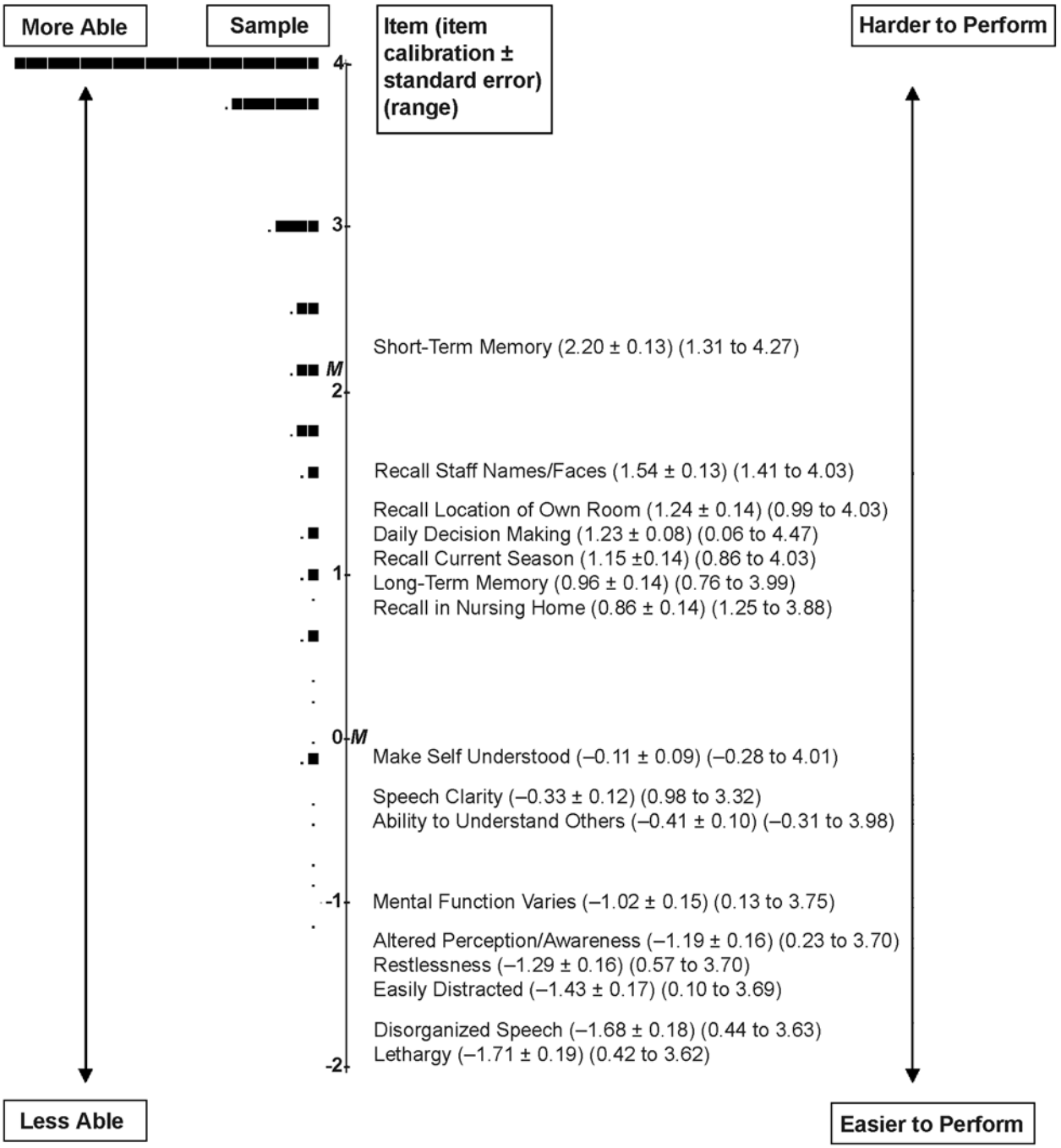

Figure 3.
Cognition subscale of Minimum Data Set: Person score distribution and item difficulty hierarchy map. Each "■" indicates 11 persons and each "." indicates up to 10 persons. *M* represents the mean of person ability measures (left) and item difficulty calibrations (right).

category misfit, with outfit MNSQ statistics slightly greater than the criterion of 2.

DIF analysis showed that few cognition items exhibited significant DIF. When comparing patients with stroke to those with amputation, we found that only two items (making self understood and speech clarity) were more challenging for the patients with stroke ($p < 0.001$). Similarly, when comparing patients with stroke to those with

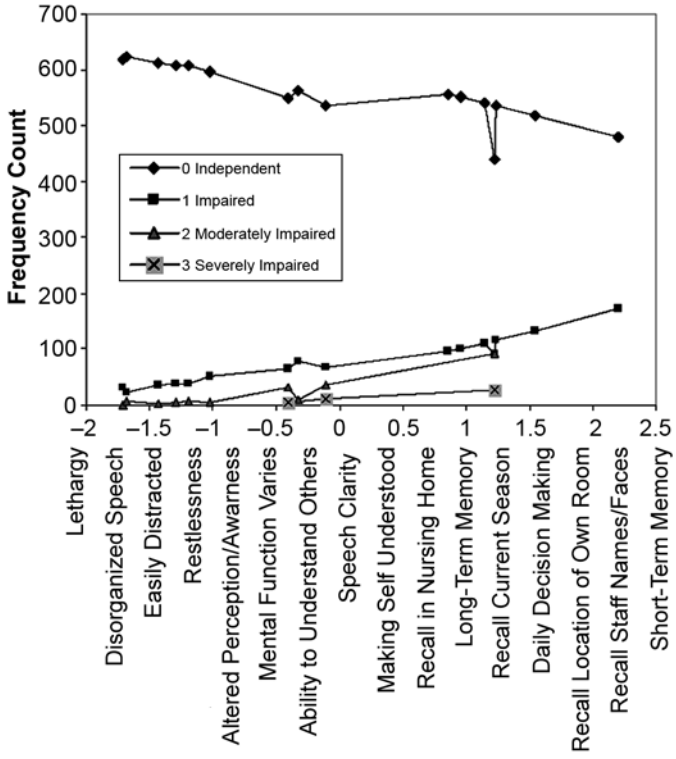

Figure 4.
Rating scale frequency count of cognition items. Items have been ordered by item difficulty from least to most difficult along *x*-axis.

orthopedic impairment, we found that the same items (making self understood and speech clarity) were more challenging for the patients with stroke ($p < 0.001$). No significant DIF was found when patients with amputation were compared with those with orthopedic impairment. While comparing patients with stroke with left hemiparesis versus stroke with right hemiparesis, we found one item (making self understood) to be more difficult for those with right hemiparesis (stroke: left brain lesion).

DISCUSSION

This study applied Rasch analysis to examine the psychometric properties of the MDS physical functioning and cognition subscales based on a sample obtained from VA's Austin Automation Center. We examined dimensionality, fit statistics, item difficulty hierarchy, targeting, separation index, rating scale usage, and DIF via the Rasch model and FA.

In general, all the items within the physical functioning and cognition subscales represented an overall functional status function, which could be further classified into separate constructs (i.e., physical functioning and cognition) in a more restrictive manner. In the physical functioning subscale, the items' difficulty levels matched well to the residents' physical conditions. These items covered a wide range of the residents' physical functioning, with the spread of items efficiently discriminating residents' performance into different physical function strata. Nonetheless, several items, including walking in corridor, walking in room, locomotion off unit, and bladder, had high fit statistics, and the additional rating of "activity did not occur during entire 7 days" had to be further investigated.

In contrast, the psychometric properties of the MDS cognition items were less sound and did not efficiently discriminate residents' performance into different cognition strata. Several items, including recall in nursing home, speech clarity, and lethargy showed high fit statistics. The score distribution of the cognition subscale was highly skewed, with 47.5 percent of the sample showing perfect scores. Similar ceiling effects in cognition subscales have been reported in several commonly used outcome measures, such as the Functional Independence Measure™ (FIM), the Outcome and Assessment Set (OASIS), the MDS for Post-Acute Care, and the Activity Measure for Post-Acute Care, for all of which more than a quarter of the sample obtained the maximum scores on the cognition subscale [39–40]. As with the present study, these ceiling effects could be a result of evaluating individuals who were not within the target groups for cognitive impairment (e.g., individuals with amputation and orthopedic impairment in the present study).

The empirical item difficulty hierarchical order of the MDS items was found to be consistent with that in the motor and cognition subscales of the FIM. In previous FIM studies, the climbing stairs, walking, and transfer-tub tasks were consistently found to be the more difficult tasks for subjects to achieve independence, whereas eating and grooming were the easiest tasks and those for which, compared with other ADLs, most patients can achieve independence [40–42]. Similar to our study, memory- and problem solving-related items were found to be more challenging than communication-related items in the FIM cognition subscale [42]. These results provide further evidence and support for the construct validity of the MDS.

In evaluating whether the response patterns of the MDS items fit the Rasch measurement model, we used the criterion of MNSQ fit statistic 1.2. The criterion was decided based on the theoretical work from Smith et al.'s

[32] simulation study and Linacre's recommendations [31] in terms of adjusting the MNSQ criterion based on the sample size to evaluate the fit of the Rasch model. More restrictive criteria are recommended as the sample size increases, especially beyond 30. Wright and Linacre once recommended MNSQ fit statistics of 0.5 to 1.7 for clinical observations [43]. Different criteria may lead to different results. In this study, we used a more restrictive criterion to evaluate the response pattern for the physical functioning and cognition items in the MDS. The finding of misfit for bladder and bowel items is not unusual and has been found elsewhere in studies related to the FIM [44–45]. Bladder and bowel items have been described as having an inherent involuntary neurological muscle control component and thus do not fit the measurement model with other consistently structured skills [46].

In this study, walking and locomotion items consistently demonstrated high fit statistics as well. These items evaluate residents' ambulatory function in two ways: (1) walking on foot and (2) locomotion, which indicates an act or the power of moving from place to place. So long as an individual has the ability to move from place A to place B, the locomotion function is not limited to movement via walking independently on foot but can also include reliance on an assistive device (such as a wheelchair). A resident with low motor function may, for example, be able to move around and go outside the facility by using a motorized wheelchair. Hence, a high score on the locomotion item is not guaranteed to be a good indicator of high motor function. Another variation, probably also the reason that the locomotion or walking items had high fit statistics, is the additional rating scale response of "activity did not occur during entire 7 days." A large portion of these ratings were given to the walking items. It is unknown, however, whether the reason for the activity not occurring during the past 7 days was because of an inability to perform the task or simply because the behavior was not observed.

Previous research with other functional assessment scales has shown different motor function hierarchies by impairment group (e.g., neurological vs musculoskeletal) in clinically logical ways but negligible DIF by sex or age [47–48]. Some investigators suggest that data should be analyzed separately for diagnostic groups if evidence of DIF across these groups is found. For instance, Tennant and colleagues adjusted their cross-cultural validity of impairment and activity limitation scales across countries [47]. Hart and colleagues developed a body-specific computerized adaptive test because of evidence of DIF by body part results [48]. In this study, the results show different motor function hierarchies by impairment group, which suggests that the data should be analyzed separately because the clinical profile is very different. However, since there is still no absolute guideline for adjusting the pairwise *t*-tests and determining the significance level of DIF analysis, more advanced DIF methods are recommended and the impact of DIF on patient outcome measures should be examined further.

There are several limitations of this study. The sample only represented individuals with stroke, amputation, and orthopedic impairment. Since this data set was from a VA database, the majority of the sample was male. Furthermore, the data selection was connected to an existing project's criteria [49]. More representative samples with larger sample sizes should be considered in future studies. While FA showed five factors with eigenvalues above 1, we performed Rasch analysis based on the two broad dimensions of physical functioning and cognition. Although the MDS shows more finely grained dimensionality, splitting items into refined dimensions would lead to small item sets in each dimension. As mentioned by Stineman et al., the challenge is to reappraise the fundamental aspects of multidimensional measurements and we should make decisions according to the measurement needed [50]. Nonetheless, further research studies should be conducted to explore whether refining the dimensions would enhance the clinical interpretation.

In this study, we applied specific rules to handle the rating scale response of 8 (activity did not occur during entire 7 days). Instead of treating 8 as missing data, we rescaled it to 4 (total dependence). This modification was based on the same reasoning in a previous study conducted by Jette et al. [40], in which they described that "the most likely explanation for the activity not occurring was that the item could not be performed." To explore the phenomenon, we conducted a frequency count of the rating scales and found that most physical functioning items were infrequently scored the rating scale response of 8, except for two walking items. We further ran the Rasch analysis while considering the rating scale response of 8 as missing data. The results showed that the item difficulty hierarchical order remained the same except for the two walking items. The item difficulty level of these two walking items dropped and became easier items, nearly at the same difficulty level as hygiene and transfer items. Our hypothesis was that in contrast to activities such as eating, hygiene, or dressing, which require staff intervention on a

daily basis if they cannot be performed independently by the residents, the walking items may be more likely to be rated as "activity did not occur" because they are not essential to maintaining residents' health (i.e., the staff do not need to assist residents to walk but they do need to assist residents to eat). Consequently, for those who cannot walk independently, even with an assistive device, their functional performance on walking tasks is more likely to be documented as "activity did not occur" instead of a dependency level. While these may be reasonable assumptions, the impact of including a specific rating for events not observed in a performance-based functional assessment should be further investigated.

In general, the item-level psychometrics of the MDS physical functioning and cognitive subscales parallel those of similar global functional scales, such as the FIM and OASIS. Physical functioning and cognition subscales appeared to form distinct constructs, with the physical function measure performing better than the cognition measure. Of particular concern is the ceiling effect of the cognition subscale. As CMS continues to revise the MDS, it is critical to evaluate physical functioning and cognition measures within the data set as to their effectiveness in monitoring patient and facility outcomes and their appropriateness for research purposes.

ACKNOWLEDGMENTS

This material was based on work supported by the University of Florida Research Opportunity Fund (grant 33070612), as well as the Health Services Research and Development and Rehabilitation Research and Development Services of the VA Rehabilitation Outcomes Research Center of Excellence, North Florida/South Georgia Veterans Health System.

The views expressed are those of the authors and do not necessarily reflect those of the VA.

The authors have declared that no competing interests exist.

REFERENCES

1. U.S. Census Bureau. Global population profile: 2002 and beyond [Internet]. Washington (DC): U.S. Census Bureau; c2008 [updated 2008 Oct 16; cited 2007 May]. About 1 screen. Available from: http://www.census.gov/ipc/www/wp02.html/.
2. Bachelder JM, Hilton CL. Implications of the Americans With Disability Act of 1990 for elderly persons. Am J Occup Ther. 1994;48(1):73–81. [PMID: 8116787]
3. Verbrugge LM, Jette AM. The disablement process. Soc Sci Med. 1994;38(1):1–14. [PMID: 8146699]
4. National Center for Health Statistics. Disabilities/limitations [Internet]. Hyattsville (MD): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; c2008 [updated 2008 Dec 17; cited 2006 Jul 12]. About 2 screens. Available from: http://www.cdc.gov/nchs/fastats/disable.htm/.
5. Pedone C, Ercolani S, Catani M, Maggio D, Ruggiero C, Quartesan R, Senin U, Mecocci P, Cherubini A; GIFA Study Group. Elderly patients with cognitive impairment have a high risk for functional decline during hospitalization: The GIFA Study. J Gerontol A Biol Sci Med Sci. 2005;60(12):1576–80. [PMID: 16424291]
6. Claesson L, Lindén T, Skoog I, Blomstrand C. Cognitive impairment after stroke—Impact on activities of daily living and costs of care for elderly people: The Göteborg 70+ Stroke Study. Cerebrovasc Dis. 2005;19(2):102–9. [PMID: 15608434]
7. National Center for Health Statistics [homepage on the Internet]. Hyattsville (MD): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; c2008 [updated 2008 Mar 13; cited 2006 Jul 24]. Nursing home care; [about 2 screens]. Available from: http://www.cdc.gov/nchs/fastats/nursingh.htm/.
8. Jones A. The National Nursing Home Survey: 1999 summary. Vital Health Stat 13. 2002;(152):1–116. [PMID: 12071118]
9. Sahyoun NR, Pratt LA, Lentzner H, Dey A, Robinson KN. The changing profile of nursing home residents: 1985–1997. Aging trends. No. 4. Hyattsville (MD): National Center for Health Statistics; 2001.
10. Morris JN, Hawes C, Fries BE, Phillips CD, Mor V, Katz S, Murphy K, Drugovich ML, Friedlob AS. Designing the national resident assessment instrument for nursing homes. Gerontologist. 1990;30(3):293–307. [PMID: 2354790]
11. Health Care Financing Administration. Minimum Data Set, 2.0. Washington (DC): U.S. Government Printing Office; 1998.
12. Casten R, Lawton MP, Parmelee PA, Kleban MH. Psychometric characteristics of the minimum data set I: Confirmatory factor analysis. J Am Geriatr Soc. 1998;46(6):726–35. [PMID: 9625189]
13. Morris JN, Nonemaker S, Murphy K, Hawes C, Fries BE, Mor V, Phillips C. A commitment to change: Revision of HCFA's RAI. J Am Geriatr Soc. 1997;45(8):1011–16. [PMID: 9256856]
14. Graney MJ, Engle VF. Stability of performance of activities of daily living using the MDS. Gerontologist. 2000; 40(5):582–86. [PMID: 11037937]

15. Lawton MP, Brody EM. Assessment of older people: Self-maintaining and instrumental activities of daily living. Gerontologist. 1969;9(3):179–86. [PMID: 5349366]
16. Snowden M, McCormick W, Russo J, Srebnik D, Comtois K, Bowen J, Teri L, Larson EB. Validity and responsiveness of the Minimum Data Set. J Am Geriatr Soc. 1999; 47(8):1000–1004. [PMID: 10443863]
17. Hawes C, Phillips CD, Mor V, Fries BE, Morris JN. MDS data should be used for research. Gerontologist. 1992; 32(4):563–64. [PMID: 1427264]
18. Teresi JA, Holmes D. Should MDS data be used for research? Gerontologist. 1992;32(2):148–49. [PMID: 1577305]
19. Lawton MP, Casten R, Parmelee PA, Van Haitsma K, Corn J, Kleban MH. Psychometric characteristics of the minimum data set II: Validity. J Am Geriatr Soc. 1998;46(6): 736–44. [PMID: 9625190]
20. Hawes C, Morris JN, Phillips CD, Mor V, Fries BE, Nonemaker S. Reliability estimates for the Minimum Data Set for nursing home resident assessment and care screening (MDS). Gerontologist. 1995;35(2):172–78. [PMID: 7750773]
21. Dallmeijer AJ, De Groot V, Roorda LD, Schepers VP, Lindeman E, Van den Berg LH, Beelen A, Dekker J; FuPro Study Group. Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: A study using Rasch analysis. J Rehabil Med. 2007;39(2):163–69. [PMID: 17351700]
22. Duncan PW, Bode RK, Min Lai S, Perera S. Glycine Antagonist in Neuroprotection Americans Investigators. Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. Arch Phys Med Rehabil. 2003;84(7): 950–63. [PMID: 12881816]
23. Franchignoni F, Giordano A, Ferriero G, Orlandini D, Amoresano A, Perucca L. Measuring mobility in people with lower limb amputation: Rasch analysis of the mobility section of the prosthesis evaluation questionnaire. J Rehabil Med. 2007;39(2):138–44. [PMID: 17351696]
24. Hsieh CL, Jang Y, Yu TY, Wang WC, Sheu CF, Wang YH. A Rasch analysis of the Frenchay Activities Index in patients with spinal cord injury. Spine. 2007;32(4):437–42. [PMID: 17304134]
25. Buchanan JL, Andres PL, Haley SM, Paddock SM, Zaslavsky AM. An assessment tool translation study. Health Care Financ Rev. 2003;24(3):45–60. [PMID: 12894634]
26. Jette AM, Haley SM, Ni P. Comparison of functional status tools used in post-acute care. Health Care Financ Rev. 2003; 24(3):13–24. [PMID: 12894632]
27. Linacre JM. A user's guide to WINSTEPS. v 3.16. Chicago (IL): MESA Press; 2005.
28. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah (NJ): Lawrence Erlbaum Associates; 2001.
29. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47(2):149–74.
30. Thurstone LL. Attitudes can be measured. Am J Sociology. 1928;33(4):529.
31. Linacre JM. Size vs. significance: Standardized chi-square fit statistic. Rasch Meas Trans. 2003;17:918.
32. Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. J Outcome Meas. 1998;2(1):66–78. [PMID: 9661732]
33. Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch Meas Trans. 2002;16(2):878.
34. Wright BD. Reliability and separation. Rasch Meas Trans. 1996;9(4):472.
35. Wright BD, Masters GN. Number of person or item strata. Rasch Meas Trans. 2002;16(3):888.
36. Wright BD, Masters GN. Rating scale analysis. Chicago (IL): MESA Press; 1982.
37. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas. 2002;3(1):85–106. [PMID: 11997586]
38. Kok FG, Mellenbergh GJ, Van Der Flier H. Detecting experimentally induced item bias using the iterative logit method. J Educ Meas. 2005;22(4):295–303.
39. Coster WJ, Haley SM, Ludlow LH, Andres PL, Ni PS. Development of an applied cognition scale to measure rehabilitation outcomes. Arch Phys Med Rehabil. 2004; 85(12):2030–35. [PMID: 15605343]
40. Jette AM, Haley SM, Ni P. Comparison of functional status tools used in post-acute care. Health Care Financ Rev. 2003; 24(3):13–24. [PMID: 12894632]
41. Velozo CA, Magalhaes LC, Pan AW, Leiter P. Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. Arch Phys Med Rehabil. 1995;76(8):705–12. [PMID: 7632124]
42. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil. 1994;75(2): 127–32. [PMID: 8311667]
43. Wright BD, Linacre JM. Reasonable mean square fit values. Rasch Meas Trans. 1994;8:370.
44. Dallmeijer AJ, Dekker J, Roorda LD, Knol DL, Van Baalen B, De Groot V, Schepers VP, Lankhorst GJ. Differential item functioning of the Functional Independence Measure in higher performing neurological patients. J Rehabil Med. 2005;37(6):346–52. [PMID: 16287665]
45. Küçükdeveci AA, Yavuzer G, Elhan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. Clin Rehabil. 2001;15(3):311–19. [PMID: 11386402]
46. Fisher WP Jr. Physical disability construct convergence across instruments: Towards a universal metric. J Outcome Meas. 1997;1(2):87–113. [PMID: 9661716]

47. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sorensen F, Marincek C, Burger H, Phillips S. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project. Med Care. 2004;42(1 Suppl):I37–48. [PMID: 14707754]
48. Hart DL, Mioduski JE, Stratford PW. Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. J Clin Epidemiol. 2005; 58(6):629–38. [PMID: 15878477]
49. Velozo CA, Byers KL, Wang YC, Joseph BR. Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. J Rehabil Res Dev. 2007;44(3):467–78. [PMID: 18247243]
50. Stineman MG, Jette A, Fiedler R, Granger C. Impairment-specific dimensions within the Functional Independence Measure. Arch Phys Med Rehabil. 1997;78(6):636–43. [PMID: 9196472]

Submitted for publication November 7, 2007. Accepted in revised form May 22, 2008.