**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

JaNae Holloway Date

Approaches to Harmonizing Two Functional Assessment Instruments in a Longitudinal Data Set of Aging Adults in the Unites States

By

JaNae Holloway, B.S.
Master of Science in Public Health
Emory University
Department of Biostatistics and Bioinformatics

---

John Hanfelt, Ph.D.
(Thesis Advisor)

---

Felicia Goldstein, Ph.D.
(Reader)

# Approaches to Harmonizing Two Functional Assessment Instruments in a Longitudinal Data Set of Aging Adults in the Unites States

By

JaNae Holloway
Bachlor of Science
University of Georgia
2019

Thesis Advisor: John Hanfelt, PhD
Reader: Felicia Goldstein, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree
of Master of Science in Public Health in
Biostatistics and Bioinformatics.
2021

# Abstract

Approaches to Harmonizing Two Functional Assessment Instruments in a Longitudinal Data Set of Aging Adults in the Unites States

By JaNae Holloway

**Background:** Declining functional ability in aging adults can be an early predictor of cognitive impairment. Individuals with mild cognitive impairment (MCI) are at a higher risk of developing dementia, which severely effects memory, language, problem solving skills, and other thinking abilities. It is essential to have a valid and reliable tool to measure functional ability. While there are multiple tools available to clinicians, there is no standardized harmonization to measure the trajectory of functional impairment over time using all available patient data.

**Materials and Methods:** We compared the use of linear regression and item response theory (IRT) in harmonizing two functional ability instruments, the Functional Assessment Questionnaire (FAQ) and the Modified ADL/IADL (MAIADL) Scale, for $n = 179$ patients in the Emory Cognitive Neurology Data Set (NeuCog). We evaluated agreement between the two measurements by calculating the concordance correlation coefficients (CCC) and evaluated model performance by calculating $R^2$ and root mean square error (RMSE). Lastly, we conducted IRT on FAQ scores from $n = 24,038$ participants in the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) and calculated new weighted FAQ scores. We performed survival analysis to clinically evaluate these new scores.

**Results:** We found that the agreement was best between factor scores (CCC=0.8585) from our IRT model compared to original scores (CCC=0.7320) and weighted scores (0.7589). Additionally, the RMSE was lowest for the linear regression model using factor scores (6.96) compared to original (17.51) and weighted (17.37) scores, however $R^2$ was highest for the model using weighted scores (0.7983). After controlling for demographic and clinical covariates, we found that FAQ significantly contributed to both time to death (HR=0.99, 95% CI: 0.99, 0.99) and time to dementia (HR=0.98, 95% CI: 0.98, 0.98), however the new weighted FAQ did not provide additional information regarding risk for time to death or time to dementia compared to the original FAQ scores.

**Conclusion:** In this paper we demonstrated the usefulness of IRT in measuring functional ability compared to the standard method of using total FAQ and MAIADL Scale scores. We believe these results can be applied in a clinical setting to assist in diagnosing patients with MCI or dementia.

# Approaches to Harmonizing Two Functional Assessment Instruments in a Longitudinal Data Set of Aging Adults in the Unites States

By

JaNae Holloway
Bachlor of Science
University of Georgia
2019

Thesis Advisor: John Hanfelt, PhD
Reader: Felicia Goldstein, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree
of Master of Science in Public Health in
Biostatistics and Bioinformatics.
2020

## Acknowledgements

I would like to thank the Department of Biostatistics and Bioinformatics at Emory University for their constant support and encouragement while pursuing my Master's degree. I would especially like to thank my thesis advisor, Dr. John Hanfelt, for granting me the opportunity to work on this project. He was invaluable to my success at Rollins School of Public Health from teaching me in the classroom to guiding me throughout this entire research experience. I would also like to thank my faculty advisor, Dr. Christina Mehta, who has continued to support and guide my decisions. She has been integral to my success at Emory through teaching and mentoring.

This project would not be possible without collaboration from the Emory Alzheimer's Disease Research Center and the National Alzheimer's Coordinating Center. I would like to thank the participants who contributed time and data to this study as well as the investigators for their support and data utilization. I especially want to thank my clinical collaborator, Dr. Felicia Goldstein, for her constant advice and feedback on my drafts, as well as Mrs. Noy Hawkins for creating my research data set and Dr. James Lah for answering all my questions regarding the clinical assessments. They have helped me gain perspective on the application of biostatistics to the real world of public health.

I would like to thank the 2021 cohort of biostatistics students at Rollins School of Public Health for contributing to my educational experience. Without our study groups, job search support systems, and long lasting friendships, I would not be as successful as a student or as a future professional.

Lastly I would like to thank my family, specifically my dad Jeff, my mom Debi, my brother Thorson, and my sister Brytta, for their unwavering support throughout my entire educational and professional career. Without them, I would not have the confidence to pursue higher education and become a biostatistician. I will forever be indebted to them and hope to make them proud.

# Contents

# 1 Introduction

Assessing functional ability is critical in clinical care of aging populations. Mild cognitive impairment (MCI), which occurs in 15% - 20% of people aged 65 years or older, is associated with declining cognitive abilities which, although noticeable, do not severely affect every day activities [1]. Individuals diagnosed with MCI are at an increased risk of developing dementia, characterized by loss of memory, language, problem-solving skills, and other thinking abilities [1,2]. Alzheimer's Disease (AD) is the most common cause of dementia and makes up 60% - 80% of cases. AD is a progressive neurodegenerative disorder caused by irreversible and progressive brain damage [2]. Unlike MCI, dementia severely interferes with normal daily life and can negatively affect behavior, feelings, and relationships. Risk factors include advanced age, genetic predisposition, and cardiovascular diseases (CVD) such as hypertension and diabetes [2]. While nonpharmacologic intervention (e.g., exercise) and pharmacological treatment (e.g., Aricept) may slow the progression of MCI and dementia, there is currently no known cure.

Compromised functional ability in individuals with dementia can be physically, economically, and socially devastating for both patients and their families. Therefore, it is important to have a valid and reliable measurement tool. Specifically, functional assessment instruments aim to measure instrumental activities of daily life (IADLs), which are complex competency skills necessary for independent living [3]. In 1969, Lawton and Brody developed the Instrumental Activities of Daily Living (IADL) Scale to assess activities which cannot be captured by only evaluating basic activities of living (ADLs) [3, 4]. A second tool used for measuring IADLs in aging populations, the Functional Activities Questionnaire (FAQ), was introduced by Pfeffer, et. al in 1982 [5]. Assessing IADLs can identify the onset of both physical and cognitive decline in older patients, allowing clinicians to guide care optimization unique to individuals' rehabilitation needs [6].

Tracking change in cognition over time requires use of repeated measure analysis, but this can be difficult when the administration of specific assessment tools changes. The Emory University Cognitive Neurology Clinics began administering the Modified Activities/Instrumental Activities of Daily Living (MAIADL) Scale, which includes a combination of items from the Lawton and Brody Physical Self-Maintenance and IADL Scales, to measure patients' functional ability. In contrast, the Alzheimer's Disease Research Centers evaluated functional ability using the FAQ. While no studies have analyzed the reliability and validity of the MAIADL Scale, the inter-rater reliability for the Lawton and Brody scales was found to be 85% and validity was established by measuring correlations between the IADL Scale and four other measurements of functional status (all $p < 0.05$) [4]. Pfeffer, et al. demonstrated the FAQ has high reliability (>90%) and

good sensitivity (85%) for determining functional impairment [5]. Additionally, it has been shown to accurately discriminate between different functional levels in older adults and is sensitive to change, supporting its value in research and clinical use [7]. In 2019, the Emory Cognitive Neurology Clinic started employing the FAQ to be consistent with the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). As a result, individuals in Emory's Neurology-Cognitive (NeuCog) data set have a combination of FAQ and MAIADL Scale scores over time. Using only visits with common tests to measure trajectory of functional impairment discards important information. Therefore, to accurately diagnose and predict rates of progression between cognitively normal, MCI, and dementia, it is critical to make use of all available data by harmonizing the FAQ and MAIADL Scale scores.

One analytic approach to harmonizing these instruments would be to create a composite score by combining the individual items. This would place the scores on a standardized scale, however each item would contribute equally to the overall score, potentially resulting in imprecise measurement of underlying functional impairment. Another approach, which would use all tests at all visits, is latent variable analysis. This method allows us to derive continuous scores representing a single latent variable on the basis of patients' responses to each test item [8, 9]. One advantage of using latent variable methods is the ability to estimate functional impairment on a common scale using different assessment tools [10,11]. Additionally, these methods address measurement error by accounting for score subsets and differing item difficulties rather than using total scores [12]. Thus, we can further explore interrelationships between specific cognitive domains.

Mental health research has increasingly applied different latent analysis methods to study psychometrics. One study suggested that using item response theory (IRT) to score the Alzheimer's Disease Assessment Scale-cognitive (ADAS-cog), an instrument which assesses cognitive change for individuals with AD, measures cognitive dysfunction more precisely than using the total score [12]. Another study evaluated the ability of factor analysis to use all available data when cognitive measures changed over time [13]. They demonstrated how latent variables better accounted for error in measuring cognitive function compared to using averages of test scores [13]. IRT also provided an effective way to harmonize cognitive measures across multiple study surveys when examining two international aging studies [14].

In this study, we aim to evaluate whether latent variable analysis methods, specifically the framework of Item Response Theory (IRT), improve our ability to harmonize the FAQ and the MAIADL Scale compared to using total scores. Our secondary aim is to illustrate the flaws in the current scoring methods for FAQ and offer an alternative method to scoring this measurement. We will clinically evaluate a new FAQ scoring

method using weights generated from IRT analysis. We hypothesize that latent variable methods will more appropriately address differences in item responses when measuring functional impairment, and by using these results to weight FAQ scores we will be able to better predict time to dementia in cognitively normal and MCI individuals.

# 2 Materials and Methods

## 2.1 Data Sources

### 2.1.1 Neurology-Cognitive Data Set

Functional assessment data on over 1,000 cognitively normal, MCI, or demented individuals were obtained from the Emory Neurology-Cognitive Data Set (NeuCog). NeuCog is comprised of patients aged 18 years or older evaluated at either the Cognitive Neurology Clinic or the Emory Alzheimer's Disease Research Center in Atlanta, Georgia. Baseline information collected includes the demographic characteristics of age, gender, race, and years of education as well as the clinical characteristics of cognitive diagnosis, Mini-Mental State Examination (MMSE), an overall measure of cognitive status, and Geriatric Depression Scale (GDS), a measure of mood. Each patient is followed longitudinally for reevaluation of their cognitive status, generally on an annual basis although this may vary depending on the clinician's preference.

In addition to the above, functional impairment is assessed at every visit either via the FAQ or MAIADL Scale. There were 179 participants in this data set with both FAQ and MAIADL measurement data available. For the purposes of harmonizing these instruments, we wanted to use the two visit dates where the different instruments were administered to be as close in time as possible. Therefore, if an individual had multiple test scores, we extracted those from visits with the closest relative dates.

### 2.1.2 Uniform Data Set

Additional functional assessment data were obtained from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) to clinically evaluate the new FAQ measurement after harmonization on a larger sample. Since 2005, 29 NIH-funded Alzheimer's Disease Centers (ADCs) across the country have contributed data on over 40,000 participants [15]. Each participant is followed longitudinally with cognitive and functional assessments obtained annually. While the administrative procedures vary across the ADCs, UDS consistently captures the same demographic and clinical variables as NeuCog with the addition of Neuropsychiatric Inventory-Questionnaire (NPI-Q), a measure of psychiatric symptoms, and Hachinski Ischemic Score, a measure of vascular disease. This study leverages NACC UDS participants who were enrolled by June 2015 and followed until the November 2019 data freeze.

Participants who were either visually or hearing impaired at baseline were excluded as this could affect their ability to perform IADLs, unrelated to functional ability. The final data set consisted of 24,038 cognitively normal, MCI, and demented patients with baseline FAQ measurement data.

## 2.2 Instrumental Activities of Daily Living Scales

The NACC adapted FAQ asks each participant if they have had difficulty in the past four weeks doing any of the following ten activities: 1) writing checks, paying bills, or balancing a checkbook; 2) assembling tax records, business affairs, or other papers; 3) shopping alone for clothes, household necessities, or groceries; 4) playing a game of skill such as bridge or chess, working on a hobby; 5) heating water, making a cup of coffee, turning off the stove; 6) preparing a balanced mean; 7) keeping track of current events; 8) paying attention to and understanding a TV program, book, or magazine; 9) remembering appointments, family occasions, holidays, medications; 10) traveling out of the neighborhood, driving, or arranging to take public transportation [5]. Each item is scored as zero (normal), one (has difficulty, but does by self), two (requires assistance), or three (dependent) (see Supplemental Table 2). If the item is not applicable to the individual or the answer is unknown, that item is disregarded in the final score. The final score is calculated as a sum of the item scores with zero representing completely functionally independent and 30 representing completely functionally dependent.

The Cognitive Neurology Clinic's MAIADL Scale combines Lawton and Brody's Physical Self-Maintenance Scale, which assess ADLs including toileting, feeding, dressing, grooming, physical ambulation, and bathing, and Lawton and Brody's IADL Scale, which assess more complex activities including ability to use the telephone, shopping, housekeeping, laundry, model of transportation, responsibility for own medications, and ability to handle finances [4]. In addition, Emory's Cognitive Neurology MAIADL Scale incorporates a question for driving, resulting in 15 total test items. While Lawton and Brody proposed each question be scored as either a zero or a one, it can be scored in various ways to fit specific assessment goals with comparable validity [16]. The Emory Cognitive Neurology Clinic employs a multi-point scoring system in which individuals begin with the highest possible score of 51, and points are subtracted depending on their level of impairment for each item (see Supplemental Table 1). Zero represents an individual who is completely functionally dependent, and 51 represents an individual who is completely functionally independent.

Interpretation of IADL scale scores are subjective. Diagnosis into the categories cognitively normal, MCI, or demented depends not only on cognitive test results but also on clinician opinion and medical history.

## 2.3 Harmonization Analyses

### 2.3.1 Linear Regression

We first calculated overall FAQ and MAIADL Scale scores by summing the individual item scores and dividing by the maximum possible points for each assessment. If a participant's response to a question was either missing, not applicable (NA), or unknown, this question did not contribute to their overall score. For example, if on the FAQ question asking whether this subject has difficulty preparing a balanced meal the answer was marked as NA, then the overall score would be calculated as a percentage out of 27 points rather than 30. This resulted in one overall score for each test measured as a percentage ranging from zero, representing completely functionally dependent, to 100, representing completely functionally independent. We ranked the scores and measured their correlation by calculating Kendall's Tau as the following:

$$\text{Kendall's Tau} = \frac{C - D}{C + D}$$

where $C$ represents the concordant pairs and $D$ represents the discordant pairs [17].

Next, we performed linear regression using SAS to assess the relationship between overall FAQ and overall MAIADL scores after adjusting for the test dates [18]. We conducted model selection by comparing the coefficient of determination ($R^2$), mean square error (MSE), and Mallow's $C_p$ for four potential models with the following covariates:

Model 1: $MAIADL$

Model 2: $MAIADL$, $MAIADL^2$

Model 3: $MAIADL$, $DATE$

Model 4: $MAIADL$, $MAIADL^2$, $DATE$

The primary model was defined as:

$$\widehat{FAQ}_i = \hat{\beta}_0 + \hat{\beta}_1 MAIADL_i + \hat{\beta}_2 MAIADL_i^2 + \hat{\beta}_3 DATE_i,\ i = 1, ..., 179$$

where $\widehat{FAQ}_i$ is the expected FAQ score for individual $i$. Covariates included the overall MAIADL Scale score for individual $i$ (as both a linear and quadratic term) and the date difference, calculated as the number of days between the calendar date when FAQ was administered and the calendar date when MAIADL Scale was administered. We centered MAIADL scores to alleviate issues of multicolinearity and confirmed covariate significance by performing partial F tests. We evaluated homoscedasticity by using the Breusch-Pagan test, assessed linearity and normality by examining diagnostic plots, and confirmed existence and independence of observations. We examined the impacts of using the transformations $FAQ^* = \ln(FAQ)$

and $FAQ^* = \sqrt{FAQ}$ on heteroscedasticity within the data.

Finally, we evaluated model performance by calculating the coefficient of determination as:

$$R^2 = 1 - \frac{\sum_{i=1}^{179} (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

and the root mean square error (RMSE) as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{179} (y_i - \hat{y}_i)^2}{179}}$$

Here $y_i$ is the observed value of FAQ for individual $i$, $\hat{y}_i$ is the predicted value of FAQ using our multiple linear regression model defined above, and $\bar{y}$ is the average observed FAQ score. Additionally, we compared the two continuous measures using Lin's Concordance Correlation Coefficient (CCC) calculated as:

$$r_c = \frac{2 r s_x s_y}{(\bar{x} - \bar{y})^2 + s_x^2 + s_y^2}$$

where $r$ is the correlation coefficient between FAQ and MAIADL Scale, $s_x$ is the standard deviation of MAIADL Scale, $s_y$ is the standard deviation of FAQ, $\bar{x}$ is the average MAIADL Scale score, and $\bar{y}$ is the average FAQ score [19].

### 2.3.2 Item Response Theory

Latent traits are unobserved traits, such as functional ability, which are reflections of observed traits, such as ability to handle finances, remembering appointments, etc. Item response theory (IRT) is a type of latent trait method where the latent variable is continuous and the observed variable is discrete. It requires the latent trait to be unidimensional, meaning it only measures one latent trait [20]. Therefore, to confirm that our data from NeuCog satisfied the assumptions for IRT, we examined the eigenvalues for each factor from factor analyses of FAQ and MAIADL Scale as well as the ratios between the first and second eigenvalues. Large ratios imply the first factor accounts for a majority of the model variance, meaning the data is unidimensional [21]. To further support our exploratory factor analyses, we performed confirmatory factor analysis using the Tucker-Lewis Index [22]. A value of 0.95 or greater represents adequate fit to the observed data with 1 being a perfect fit.

Because the test questions employ ordinal scoring methods, we used a graded response model with observed items $x_1, ..., x_p$ and unobserved latent traits for each individual $y_1, ..., y_q \overset{iid}{\sim} \text{Normal}(0, 1)$ where $q = 1, ..., 179$

[23]. Here latent factors represent a participant's functional ability. We assume local independence such that $f(x_1, ..., x_p|y_1, ..., y_q) = \prod_{j=1}^{p} f(x_j|y_1, ..., y_q)$. The proportional odds model for each level $c$ of ordinal item $j$ is given by:

$$\log \frac{P(X_j \leq c|y_1, ..., y_q)}{P(X_j > c|y_1, ..., y_q)} = \alpha_{j(c)} + \alpha_{j1}y_1 + ... + \alpha_{jq}y_q, \quad c \in \{1, 2, ..., m_j - 1\}$$

or equivalently,

$$P(X_j = c|y_1, ..., y_q) = \text{expit}\left(\alpha_{j(c)} + \alpha_{j1}y_1 + ... + \alpha_{jq}y_q\right) - \text{expit}\left(\alpha_{j(c-1)} + \alpha_{j1}y_1 + ... + \alpha_{jq}y_q\right)$$

where for any constant $c$, $\text{expit}(c) = \frac{\exp(c)}{1+\exp(c)}$ with c varying with each question. Our model measures the probability that an individual with functional ability $y$ receives a score $c$. The threshold parameters $\alpha_{j(c)}$ describe how difficult it is for a typical individual in the study population to achieve a score of $c$ or lower on item $j$ where $j = 1, ..., 10$ for FAQ and $j = 1, ..., 15$ for MAIADL Scale. A higher threshold value represents a difficult question whereas a lower threshold value represents an easy question. These are ordered across the levels of ordinal items such that $\alpha_{j(1)} \leq \alpha_{j(2)} \leq ... \leq \alpha_{j(m_j-1)}$. The discrimination parameters $\alpha_{jk}$ describe how well item $j$ differentiates between individuals with low functional ability and individuals with high functional ability. A higher discrimination value represents a better ability to distinguish between levels of functional impairment.

We built two separate models, one with FAQ and one with MAIADL Scale, to measure functional ability using IRT procedures in SAS [24]. We compared the item specific discrimination estimates within each scale and plotted item characteristic curves (ICCs) to determine how daily activities contributed differently to measuring overall functional ability. We obtained estimated factor scores, which measure functional ability, from each model and evaluated their relationship using CCC. To investigate the effect of time on harmonization, we calculated CCC separately for individuals with less than five years between their two assessments and for individuals with less than three years between their two assessments.

Lastly, we used the item discrimination estimates as weights for each item score. We recalculated the overall FAQ and MAIADL Scale scores using these new weighted item scores and calculated the CCC for comparison. We performed model selection by comparing $R^2$, MSE, and Mallow's $C_p$ and fit a linear regression model predicting FAQ score from MAIADL Scale score after adjusting for date. We repeated this process using the raw factor scores. We compared all three models (original scores, weighted scores, and factor scores) by

examining scatter plots and calculating $R^2$ and root mean square error (RMSE).

### 2.4 Survival Analysis

To further explore how test items differentially measure functional ability, we conducted IRT analysis for the FAQ scale using the $n = 24,038$ patients from NACC UDS. We first performed exploratory factor analysis and examined the eigenvalues to determine if IRT was appropriate for the data. Then, from the IRT model results, we again used the discrimination estimates as weights for each item and recalculated the overall FAQ score, re-scaling it to be a percentage ranging from zero to 100. We fit two multivariable Cox proportional hazards model in R - one using the original FAQ score and one using the weighted FAQ score - to assess how well FAQ predicted time to death for cognitively normal, MCI, or demented patients [25]. R employed listwise deletion to remove cases with missing data. We adjusted for clinically relevant covariates in our model including age, categorized as 0-64 years, 65-69 years, 70-74 years, 75-79 years, and $\geq$ 80 years; sex, categorized as female or male; race, categorized as black or other; education, categorized as having less than or equal to a high school degree or more than a high school degree; cognitive diagnosis categorized as normal, MCI, or demented; MMSE score captured continuously; GDS, categorized as depressed or not depressed; NPI captured continuously; and Hachinski Ischemic Scale, categorized as being at risk for vascular disease or not. The model was define as:

$$h(t) = h_0(t)e^{x^T\beta}$$

where $t$ is the follow-up time in years after the initial visit, $h_0(t)$ is the baseline hazard function, $x$ is the vector of covariates, and $\beta$ is the vector of log hazard ratios [32]. We calculated hazard ratios and 95% confidence intervals for each covariate and compared the hazard ratio for the weighted FAQ score to that of the original FAQ score.

Next we evaluated the ability for FAQ to predict time to dementia after adjusting for clinical covariates in individuals who were either cognitively normal or diagnosed with MCI at baseline. We fit two Cox proportional hazards models and again compared the hazard ratios for the original and weighted FAQ scores. We checked the proportional hazards assumption for all four models graphically.

## 3 Results

### 3.1 Patient Demographics

The full sample included 24,217 adults diagnosed as either cognitively normal (n=9,941), MCI (n=5,361), or demented (n=8,915) at baseline (Table 1). The majority of patients in the NeuCog data set were diagnosed

with dementia (63%) whereas UDS comprised mostly of cognitively normal patients (41%). Longitudinal data were available from both studies. The difference between FAQ and MAIADL Scale test dates in NeuCog ranged from zero to 10 years with a median of 104 days. The median follow-up time for participants in UDS was 3.5 years (Q1-Q3: 1.1-6.4). Participants had a median age of 73 years (Q1-Q3: 65.79). Most individuals from NeuCog were in the age group 65-69 years old (29%) whereas most individuals from UDS were $\geq$ 80 years old (24%). The NeuCog sample was 46% female, 13% black, and 13% depressed with 15% having less than or equal to a high school education compared to the UDS sample which was 58% female, 14% black, and 15% depressed with 33% having less than or equal to a high school education. Additionally 5% had were at risk for cerebrovascular disease. The median MMSE scores were 27 (Q1-Q3: 25-29) and 28 (23-29) for NeuCog and UDS, respectively. The median number of neuropsychiatric features for participants in UDS was 1 (Q1-Q3: 0-3).

There were $n = 179$ participants in NeuCog who had both FAQ and MAIADL scores available. The median score for FAQ was 70.4% (Q1-Q3: 26%-96.7%) whereas the median score for MAIADL Scale was 88.2% (Q1-3: 54.9%-100%) (Figure 1). There were 37 participants with a perfect 100% score on FAQ, and 53 participants with a perfect 100% score on MAIADL Scale. We calculated Kendall's Tau as 0.66, which implies the agreement between these two instruments is not great.

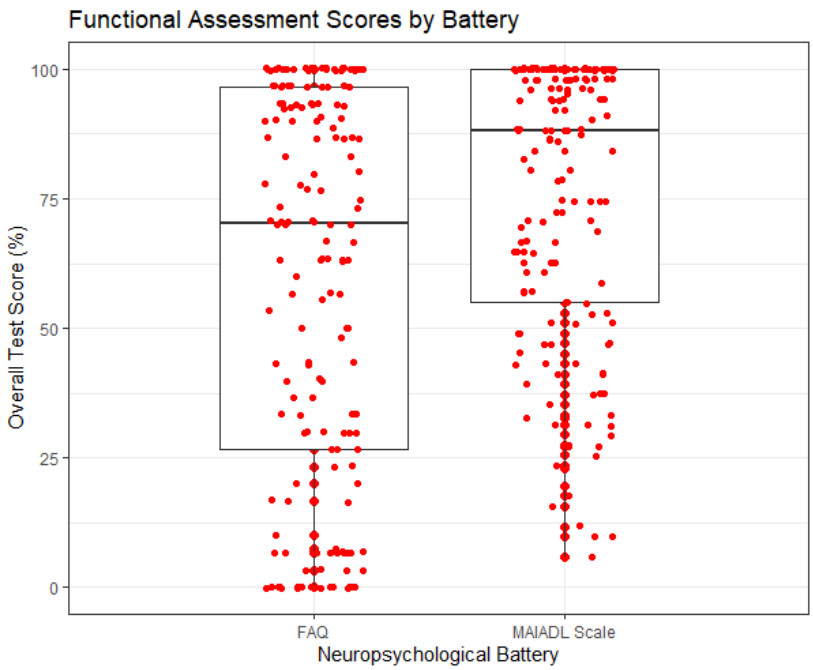


Figure 1: Distribution of overall functional assessment scores for n=179 individuals to compare Functional Assessment Questionnaire and Modified ADL/IADL Scale after scaling to be between 0% and 100%, but before applying weights from item response theory

**Table 1. Baseline Demographic and Clinical Characteristics of Study Samples**

| **Characteristic, n(%) or median (Q1-Q3)** | **Full Sample (n=24,217)** | **NeuCog Data Set (n=179)** | **NACC UDS (n=24,038)** |
|---|---|---|---|
| Age, years | 73 (65-79) | 69.4 (65.1-76.2) | 73 (65-79) |
| Age Group | | | |
| [0,65) years | 5,455 (23) | 43 (24) | 5,412 (23) |
| [65,70) years | 4,012 (17) | 52 (29) | 3,960 (16) |
| [70,75) years | 4,472 (18) | 34 (19) | 4,438 (18) |
| [75,80) years | 4,531 (19) | 27 (15) | 4,504 (19) |
| $\geq$ 80 years | 5,747 (24) | 23 (13) | 5,724 (24) |
| Sex, female | 14,019 (58) | 83 (46) | 13,936 (58) |
| Race, black | 3,348 (14) | 23 (13) | 3,325 (14) |
| Years of Education | 16 (12-18) | 16 (15-18) | 16 (12-18) |
| Education | | | |
| $\leq$ HS | 8,080 (33) | 26 (15) | 7,912 (33) |
| $>$ HS | 18,137 (67) | 152 (85) | 15,984 (67) |
| Cognitive Diagnosis | | | |
| Normal | 9,941 (41) | 34 (19) | 9,907 (41) |
| MCI | 5,361 (22) | 32 (18) | 5,329 (22) |
| Demented | 8,915 (37) | 113 (63) | 8,802 (37) |
| MMSE Score^ | 28 (23-29) | 27 (25-29) | 28 (23-29) |
| Depressed† | 3453 (15) | 20 (13) | 3,433 (15) |
| NPI | NA | NA | 1 (0-3) |
| Cerebrovascular Disease Indicator‡ | NA | NA | 1,232 (5.1) |
| FAQ Score¶ | 93.3 (56.7-100) | 70.4 (26.7-96.7) | 93.3 (57.1-100) |
| Modified ADL/IADL Scale Score¶ | NA | 88.2 (54.9-100) | NA |
| Follow-Up Time, years | NA | NA | 3.5 (1.1-6.4) |

Abbreviations: ADL=Activities of Daily Living; FAQ= Functional Assessment Questionnaire; HS = high school; IADL=Instrumental Activities of Daily Life; MCI = Mild Cognitive Impairment; MMSE = Mini-Mental State Examination; NPI = Neuropsychiatric Inventory

*Data missing for the following: MMSE Score (n=40); Depressed (n=1245); FAQ (n=473)

^Range 0-30, with more severe cognitive impairment represented by a lower score

† Determined using Geriatric Depression Scale (GDS) $\geq$ 5

‡Determined using Hachinski Ischemic Scale $\geq$ 4

¶Derived proportional score from raw scores

%: Column percents may not total to 100 due to rounding

## 3.2 Linear Regression

The overall FAQ and MAIADL Scale scores we calculated are plotted in Figure 2 below. Because there was high multicolinearity between the $MAIADL$ and $MAIADL^2$ terms ($r = 0.9848$, $p < 0.0001$) in the multiple linear regression model, we refit the model with centered MAIADL Scale scores. Taken together,

all three covariates contribute significantly to predicting overall FAQ score ($F = 220.77$, $p < 0.001$). Tests of the regression coefficients confirm that the terms $MAIADL$ ($t = 21.32$, $p < 0.0001$), $MAIADL^2$ ($t = 4.89$, $p < 0.0001$), and $DATE$ ($t = 2.08$, $p = 0.0393$) are important for predicting a FAQ. While the assumptions for linearity, normality, existence, and independence were met, the residual plots showed uneven distribution residuals and the partial plots showed fanning patterns. We confirmed there was not homogeneity of variance based on the Breusch-Pagan test for heteroscedasticity ($\chi^2 = 6.45$, $p = 0.0111$). Transforming our outcome variable, FAQ, did not alleviate this issue.

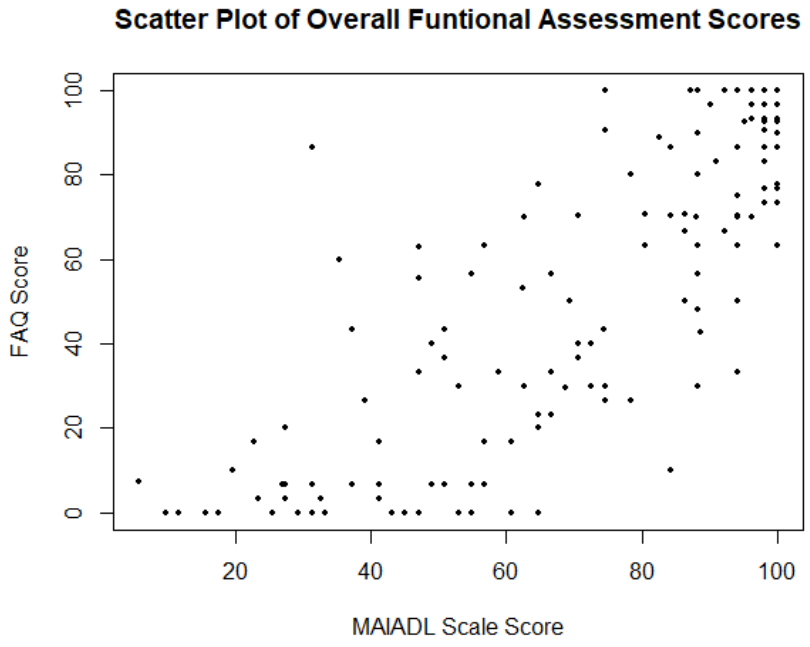


Figure 2: Scatter plot of overall Functional Assessment Questionnaire and Modified ADL/IADL Scale scores showing a slightly positive linear relationship between the two instruments with large floor and ceiling effects

We found that 79.38% of the variability in FAQ can be explained by our multiple linear regression, the root mean square error is 17.51, and there is moderate concordance between these two measurements for cognitive impairment ($r_c = 0.7320$). However, our residual plots imply linear regression does not fully capture the relationship between FAQ and MAIADL Scale scores.

### 3.3 Item Response Theory

Exploratory factor analysis of FAQ yielded a first eigenvalue of 9.05, accounting for 90% of the variance, and factor analysis of MAIADL Scale yielded a first eigenvalue of 12.95, accounting for 86% of the variance. The eigenvalues for all other factors were less than one and did not account for a significant proportion of the variance. In addition, the ratios between the first and second eigenvalues were 32.32 and 20.23 for the FAQ

and MAIADL Scale, respectively, which are considered very large. These preliminary results were further supported by confirmatory factor analysis (Tucker-Lewis Index for FAQ=0.94; for MAIADL Scale=0.96). Therefore, we can be sure we are only measuring one latent trait, functional ability, and can fit a unidimensional graded response model.

The IRT model parameter estimates for FAQ and MAIADL Scale are listed in Tables 2 and 3, respectively. In the IRT model for FAQ, the bills ($\alpha_1 = 17.36$) and taxes ($\alpha_2 = 18.21$) items have the highest ability to distinguish between low functioning and high functioning individuals while the stove ($\alpha_5 = 3.94$) and pay attention ($\alpha_8 = 3.42$) items have the lowest distinguishing ability. In the IRT model for MAIADL Scale, the shopping ($\alpha_2 = 7.23$) and food preparation ($\alpha_3 = 6.82$) items have the highest ability to distinguish between low and high functioning individuals while the toileting ($\alpha_{14} = 2.66$) and walking ($\alpha_{15} = 2.85$) items have the lowest distinguishing ability. Both models show that individual items contribute differently to measuring functional ability, and therefore should receive different weights when calculating the overall scores for these instruments.

The threshold estimates measure how difficult it is to score a $c$ or lower for a person with average functional ability where $c \in \{1, 2, 3\}$ for FAQ and varies by question for MAIADL Scale. For FAQ, it is easiest to score one or less on the remember dates item ($\alpha_{9(1)} = -0.53$) and most difficult to score a one or less on the stove item ($\alpha_{5(1)} = 0.21$). For MAIADL Scale, it is easiest to score one or less on the ability to handle finances item ($\alpha_{8(1)} = -0.21$) and most difficult to score a one or less on the feeding item ($\alpha_{12(1)} = 1.10$).

**Table 2. IRT Model Parameter Estimates for Functional Assessment Questionnaire Using NeuCog**

| Item | Discrimination Estimate | Threshold Estimates | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Bills | 17.36 | -0.23 | -0.02 | 0.20 |
| Taxes | 18.21 | -0.25 | -0.08 | 0.14 |
| Shopping | 8.17 | -0.12 | 0.12 | 0.53 |
| Games | 4.29 | 0.08 | 0.43 | 0.96 |
| Stove | 3.94 | 0.21 | 0.48 | 0.78 |
| Meal Preparation | 5.03 | 0.11 | 0.42 | 0.69 |
| Events | 4.20 | -0.03 | 0.43 | 0.83 |
| Pay Attention | 3.42 | -0.001 | 0.69 | 1.29 |
| Remember Dates | 4.56 | -0.53 | -0.16 | 0.47 |
| Travel | 5.29 | -0.30 | -0.06 | 0.23 |

**Table 3. IRT Model Parameter Estimates for Modified ADL/IADL Scale Using NeuCog**

| Item | Discrimination Estimate | Threshold Estimates | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 3A. Physical Self-Maintenance Scale | | | | | |
| Bathing | 3.39 | 1.08 | 1.55 | 1.69 | 2.20 |
| Dressing | 4.80 | 0.65 | 0.97 | 1.31 | 2.11 |
| Feeding | 3.88 | 1.10 | 1.75 | 2.14 | 2.72 |
| Grooming | 4.66 | 0.66 | 1.01 | 1.55 | |
| Toileting | 2.66 | 0.91 | 1.40 | 1.47 | 2.12 |
| Walking | 2.85 | 0.30 | 1.14 | 2.41 | 2.93 |
| 3B. Instrumental Activities of Daily Living Scale | | | | | |
| Ability to Use Telephone | 4.95 | 0.27 | 0.69 | 1.30 | — |
| Shopping | 7.23 | -0.06 | 0.33 | 1.21 | — |
| Food Preparation | 6.82 | 0.01 | 0.29 | 0.54 | — |
| Housekeeping | 5.50 | -0.02 | 0.64 | 0.82 | 1.05 |
| Laundry | 4.88 | 0.38 | 0.63 | — | — |
| Mode of Transportation | 3.99 | 0.19 | 0.30 | 0.39 | 1.78 |
| Responsible for Own Medication | 4.58 | 0.11 | 0.62 | — | — |
| Ability to Handle Finances | 5.79 | -0.21 | 0.59 | — | — |
| Driving | 4.41 | -0.12 | -0.05 | 0.17 | 0.31 |

We can further examine each test item by looking at the item characteristic curves (ICCs) (Figures 4 and 5). The lines represent the relationship between functional ability (on the x-axis) and probability of achieving a score of $c$ (on y-axis). The points at $x = 0$ describe the probability of scoring $c$ for someone who is functionally normal, i.e. their factor score is zero. The color of the line represents the specific item score $c$. For example, the blue line in the first plot of Figure 4 depicts the probability of scoring a zero on the FAQ bills item for individuals with different levels of functional ability. The FAQ curves are steepest for the bills and taxes items, further confirming these items are better at differentiating between people with low versus high functional ability. Similarly, the MAIADL Scale curves are steepest for shopping and food preparation.

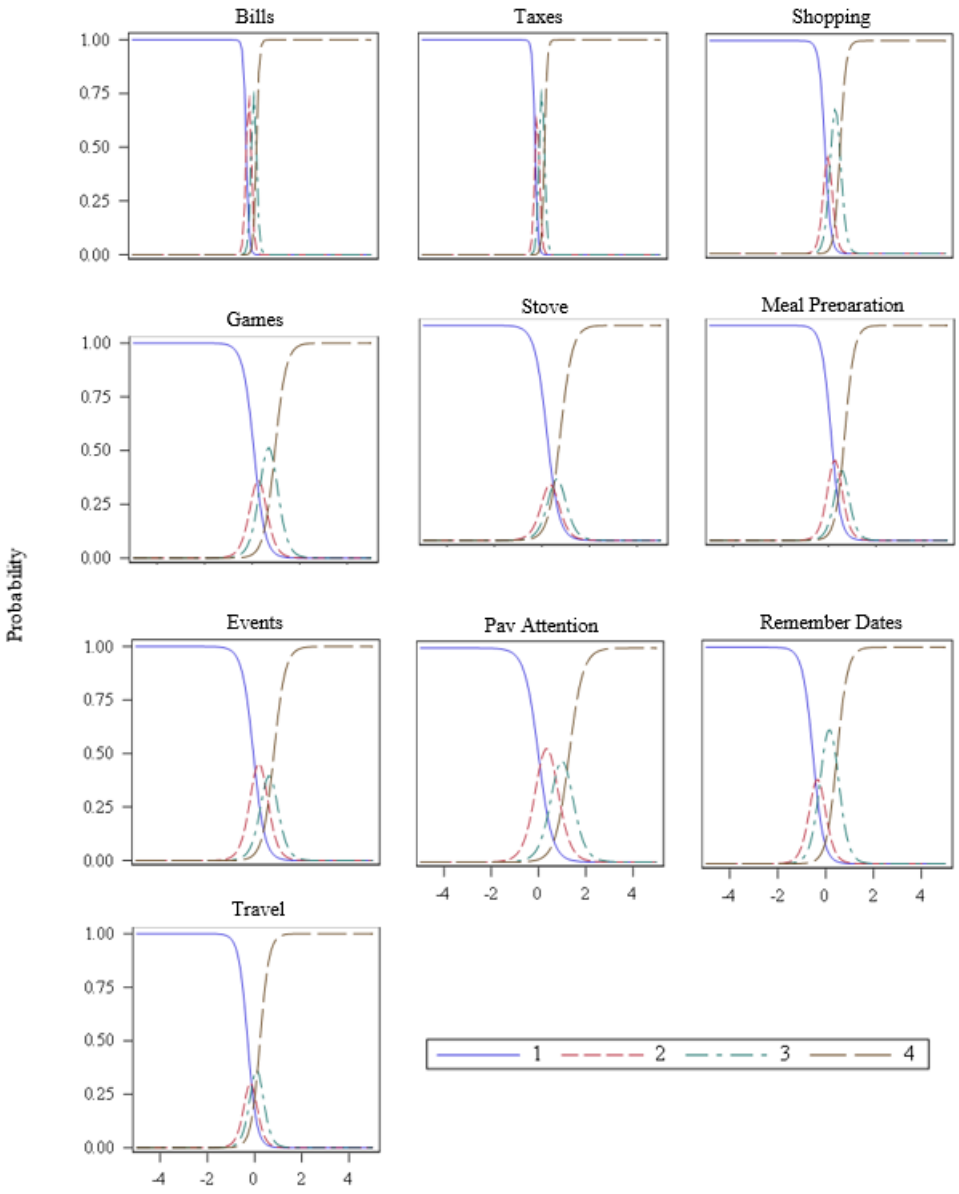


Figure 3: Item characteristic curves depicting how FAQ items measure cognitive ability differently

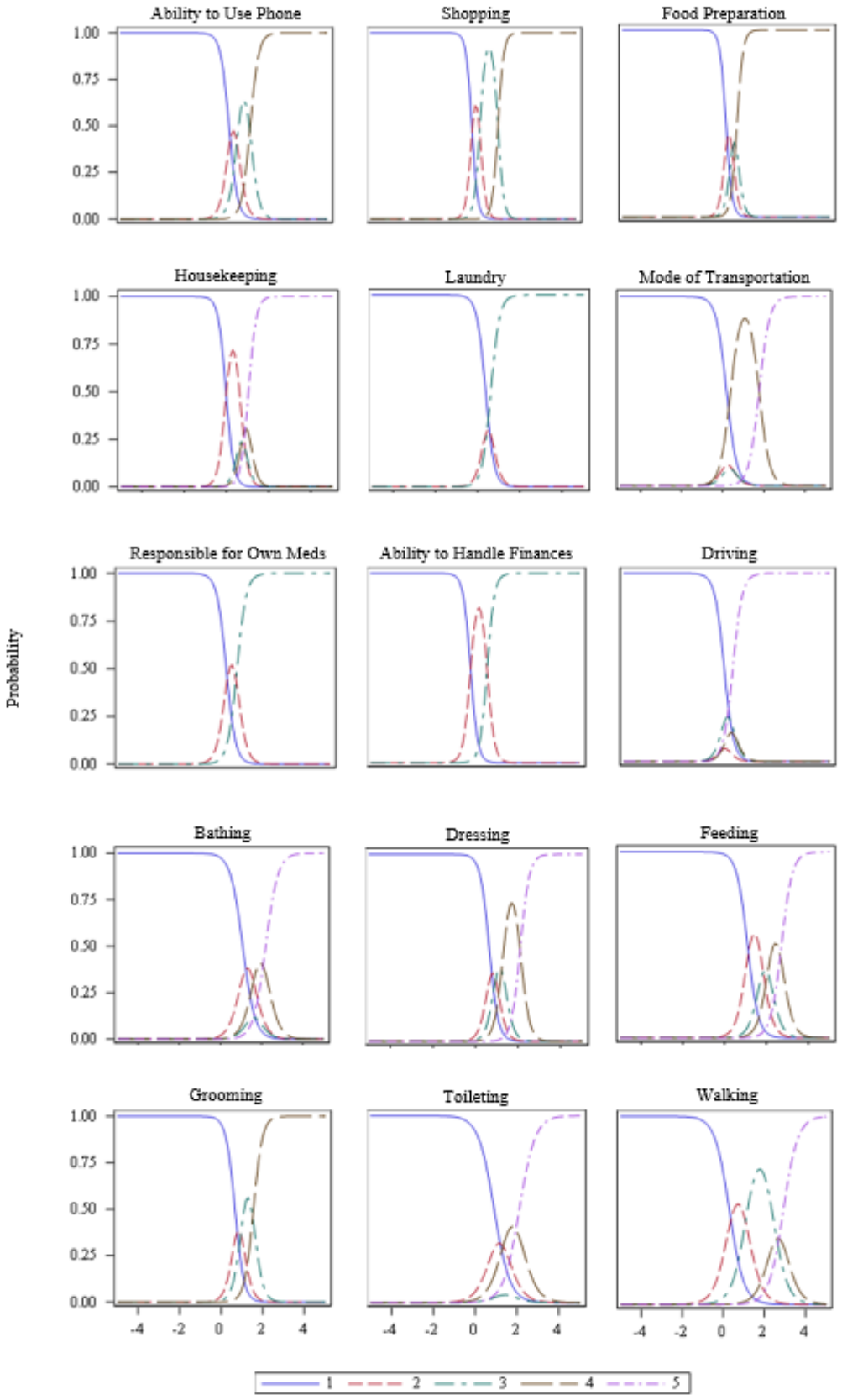


Figure 4: Item characteristic curves depicting how MAIADL Scale items measure cognitive ability differently

Figure 5A plots the factor scores from our graded response models and Figure 5B plots the new overall scores after using the discrimination estimates as weights for each item. Three individuals with more than five years between their two assessments are shown in blue and eight individuals with three to five years between their two assessments are shown in red. The CCCs between FAQ and MAIADL Scale factor scores are $r_c = 0.8484$ for all individuals, $r_c = 0.8541$ for individuals with less than five years between assessments, and $r_c = 0.8614$ for individuals with less than three years between assessments. The CCCs between FAQ and MAIADL Scale overall weighted scores are $r_c = 0.7589$ for all individuals, $r_c = 0.7615$ for individuals with less than five years between assessments, and $r_c = 0.7524$ for individuals with less than three years between assessments. We can see from the plots that overall weighted scores deviate more severely from the 45° line, resulting in lower CCCs.

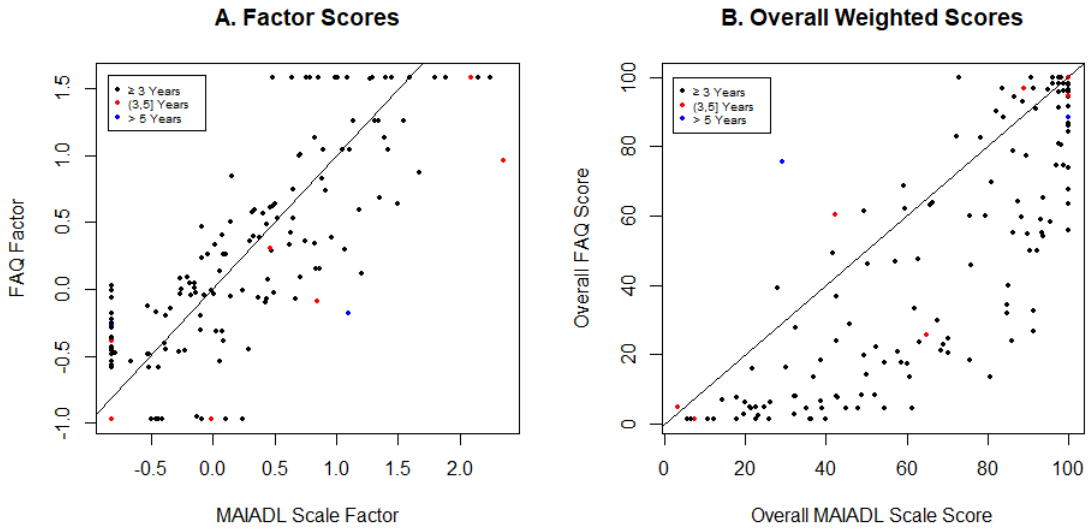


Figure 5: Scatter plots of FAQ versus MAIADL Scale overlaid with the $y = x$ line show the line of best fit is (a) linear for factor scores and (b) quadratic for overall weighted scores.

Model selection for weighted scores chose the multiple linear regression with the $MAIADL$ ($t = 22.10$, $p < 0.0001$), $MAIADL^2$ ($t = 5.40$, $p < 0.0001$), and $DATE$ ($t = 3.56$, $p = 0.0005$) terms. Univariate analyses showed that each covariate has a significant relationship with weighted FAQ score. Taken together, all three covariates contribute significantly to predicting overall weighted FAQ score ($F = 230.86$, $p < 0.0001$). We found that 79.83% of the variability in weighted FAQ can be explained by the model, and the root mean square error is 17.37.

Model selection for the factor scores chose the linear regression model with $MAIADL$ ($t = 22.12$, $p < 0.0001$)

and $DATE$ ($t = -3.45$, $p = 0.0007$) covariates. Taken together, both predictors contribute significantly to predicting FAQ factor score ($F = 250.57$, $p < 0.0001$). We found that 74.01% of the variability in FAQ factor score can be explained by the model, and the root mean square error is 0.4176. Similar to our previous linear regression model with the original FAQ and MAIADL Scale scores, both of the models for weighted scores and factor scores did not satisfy the assumption for homoscedasticity, and transforming the outcome variable did not fix this.

## 3.4 Comparison of Harmonization Techniques

The agreement between two measures of functional ability is highest for factor scores ($r_c = 0.8585$), followed by weighted total scores ($r_c = 0.7589$), then the original total scores ($r_c = 0.7320$) (Table 4). After fitting the line of best fit to the data, the model using weighted FAQ and MAIADL Scale scores has the highest $R^2$ (0.7983), and the model using factor scores representing functional ability has the lowest RMSE (6.96).

**Table 4. Comparison of Different FAQ and MAIADL Scale Scores**

| | **Original Scores** | **Weighted Scores** | **Factor Scores** |
|---|---|---|---|
| **Measure of Agreement** | | | |
| Concordance Correlation Coefficient | 0.7320 | 0.7589 | 0.8585 |
| **Model Evaluation** | | | |
| $R^2$ | 0.7938 | 0.7983 | 0.7401 |
| Root Mean Square Error | 17.51 | 17.37 | 6.96* |

*Calculated by multiplying original model RMSE (0.4176) by 100/6 to put on the same scale

## 3.5 Clinical Evaluation of Weighted FAQ Score

Exploratory factor analysis of FAQ from UDS yielded a first eigenvalue of 8.99, accounting for 90% of the variance. The eigenvalues for all other factors were less than one and did not account for a significant proportion of the variance. The ratio between the first and second eigenvalue is 37.6, which is very large. Additionally, the Tucker-Lewis Index is 0.96. Based on results from exploratory and confirmatory factor analyses, we can be confident our data is unidimensional, and it is appropriate to fit a graded response model. Table 5 displays the discrimination estimates, which were used to calculated the weighted FAQ scores, and the threshold estimates based on $N = 24,038$ participants in UDS.

**Table 5. IRT Model Parameter Estimates for Functional Assessment Questionnaire Using Using the NACC Uniform Data Set**

| Item | Discrimination Estimate | Threshold Estimates | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Bills | 6.93 | 0.24 | 0.54 | 0.85 |
| Taxes | 7.46 | 0.19 | 0.44 | 0.75 |
| Shopping | 6.35 | 0.44 | 0.78 | 1.21 |
| Games | 4.54 | 0.54 | 1.02 | 1.43 |
| Stove | 4.15 | 0.84 | 1.23 | 1.56 |
| Meal Preparation | 5.68 | 0.51 | 0.83 | 1.17 |
| Events | 4.46 | 0.46 | 0.96 | 1.46 |
| Pay Attention | 3.71 | 0.57 | 1.20 | 1.81 |
| Remember Dates | 4.74 | 0.11 | 0.60 | 1.18 |
| Travel | 5.21 | 0.28 | 0.64 | 0.94 |

Overall there were 6,065 deaths and 2,444 new dementia diagnoses throughout this study period. Table 6 displays the results from our multivariable Cox Proportional Hazards models predicting time to death for $N = 22,419$ patients using (1) the original FAQ scores and (2) the weighted FAQ scores, adjusting for demographic and clinical covariates. In both models the risk of death was significantly lower for black participants and participants who were not depressed at baseline. Compared to participants less than 65 years old, the risk of death was significantly higher for participants 70-74 years old, 75-79 years old, and $\geq$ 80 years old. Compared to cognitively normal patients, the risk of death was significantly higher for MCI patients and demented patients. Additionally, the risk of death was higher for males and those at risk for cerebrovascular disease. The risk associated with a one unit increase in MMSE Score was 0.96 (95% CI: 0.95, 0.97) and a one unit increase in NPI-Q was 1.05 (95% CI: 1.04, 1.06). For both the model using original FAQ scores and the model using weighted FAQ scores, the risk associated with a 10 point increase in the respective FAQ score was 0.94 (95% CI: 0.93, 0.95).

Table 7 displays the results from our multivariable Cox Proportional Hazards models predicting time to dementia for $N = 14,568$ patients using (1) the original FAQ scores and (2) the weighted FAQ scores, adjusting for demographic and clinical covariates. Similar to our previous models, here the risk of dementia was significantly lower for black participants and participants who were not depressed at baseline. Compared to participants less than 65 years old, the risk of dementia was significantly higher for participants 65-69 years old, 70-74 years old, 75-79 years old, and $\geq$ 80 years old. Additionally, the risk of dementia was higher for MCI patients compared to cognitively normal patients. The risk associated with a one unit increase in MMSE Score was 0.86 (95% CI: 0.84, 0.87) and a one unit increase in NPI-Q was 1.10 (95%

CI: 1.08, 1.12). The risk associated with a 10 point increase in original FAQ score was 0.81 (95% CI: 0.81, 0.82) and the risk associated with a 10 point increase in the weighted FAQ score was 0.82 (95% CI: 0.82, 0.83).

**Table 6. Multivariable Cox Proportional Hazards Model Predicting Time to Death for N=22,419 Patients at Baseline**

| | | Original FAQ | | Weighted FAQ | |
|---|---|---|---|---|---|
| **Covariate** | **Level** | **Hazard Ratio (95% CI)** | **p-value** | **Hazard Ratio (95% CI)** | **p-value** |
| Age | [0,65) years | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | [65,70) years | 1.01 (0.91, 1.12) | | 1.01 (0.91, 1.12) | |
| | [70,75) years | 1.33 (1.21, 1.46) | | 1.33 (1.21, 1.46) | |
| | [75,80) years | 1.57 (1.44, 1.71) | | 1.57 (1.44, 1.71) | |
| | $\geq$ 80 years | 2.79 (2.58, 3.02) | | 2.80 (2.59, 3.03) | |
| Sex | Female | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | Male | 1.47 (1.39, 1.55) | | 1.47 (1.39, 1.54) | |
| Race | Else | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | Black | 0.79 (0.72, 0.86) | | 0.79 (0.72, 0.86) | |
| Education | $\leq$ High School | ref | 0.3380 | ref | 0.3220 |
| | $>$ High School | 1.03 (0.97, 1.09) | | 1.03 (0.97, 1.09) | |
| Cognitive Diagnosis | Normal | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | MCI | 1.86 (1.72, 2.02) | | 1.88 (1.73, 2.04) | |
| | Demented | 2.67 (2.43, 2.93) | | 2.73 (2.49, 3.00) | |
| MMSE Score | | 0.96 (0.95, 0.97) | $< 0.0001$ | 0.96 (0.95, 0.97) | $< 0.0001$ |
| Geriatric Depression Scale | Depressed | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | Not Depressed | 0.75 (0.70, 0.80) | | 0.75 (0.70, 0.80) | |
| NPI-Q | | 1.05 (1.04, 1.06) | $< 0.0001$ | 1.05 (1.04, 1.06) | $< 0.0001$ |
| Cerebrovascular Disease | Absent | ref | $< 0.0001$ | ref | $< 0.0001$ |
| Indicator | Present | 1.55 (1.41, 1.71) | | 1.56 (1.42, 1.71) | |
| FAQ Score | | 0.99 (0.99, 0.99) | $< 0.0001$ | 0.99 (0.99, 0.99) | $< 0.0001$ |

**Table 7. Multivariable Cox Proportional Hazards Model Predicting Time to Dementia for N=14,568 Patients without Dementia at Baseline**

| | | Original FAQ | | Weighted FAQ | |
|---|---|---|---|---|---|
| **Covariate** | **Level** | **Hazard Ratio (95% CI)** | **p-value** | **Hazard Ratio (95% CI)** | **p-value** |
| Age | [0,65) years | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | [65,70) years | 1.35 (1.14, 1.59) | | 1.34 (1.14, 1.58) | |
| | [70,75) years | 1.73 (1.49, 2.02) | | 1.73 (1.48, 2.01) | |
| | [75,80) years | 2.23 (1.92, 2.59) | | 2.23 (1.92, 2.58) | |
| | $\geq$ 80 years | 2.79 (2.42, 3.22) | | 2.78 (2.41, 3.21) | |
| Sex | Female | ref | 0.8884 | ref | 0.8878 |
| | Male | 0.99 (0.92, 1.08) | | 0.99 (0.92, 1.08) | |
| Race | Else | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | Black | 0.63 (0.55, 0.72) | | 0.63 (0.55, 0.73) | |
| Education | $\leq$ High School | ref | 0.0516 | ref | $< 0.0001$ |
| | $>$ High School | 1.10 (0.10, 1.20) | | 1.10 (1.00, 1.20) | |
| Cognitive Diagnosis | Normal | ref | $< 0.0001$ | ref | $< 0.0001$ |
| | MCI | 5.55 (5.02, 6.15) | | 5.56 (5.02, 6.15) | |
| MMSE Score | | 0.86 (0.84, 0.87) | $< 0.0001$ | 0.86 (0.84, 0.87) | $< 0.0001$ |
| Geriatric Depression Scale | Depressed | ref | 0.0072 | ref | 0.0071 |
| | Not Depressed | 0.85 (0.75, 0.96) | | 0.85 (0.75, 0.96) | |
| NPI-Q | | 1.10 (1.08, 1.12) | $< 0.0001$ | 1.10 (1.08, 1.12) | $< 0.0001$ |
| Cerebrovascular Disease | Absent | ref | 0.4750 | ref | 0.4694 |
| Indicator | Present | 0.93 (0.78, 1.13) | | 0.93 (0.78, 1.12) | |
| FAQ Score | | 0.98 (0.98, 0.98) | $< 0.0001$ | 0.98 (0.98, 0.98) | $< 0.0001$ |

# 4 Discussion

In the NeuCog data set consisting of 179 participants, the majority of which were non-black males between the ages of 65 and 69 years old diagnosed with dementia and having more than a high school education, there was poor agreement between the FAQ and MAIADL Scale scores (Kendall's Tau=0.66). As a result, we sought different ways to harmonize these two instruments which both measure functional ability. Because the MAIADL Scale was an instrument designed specifically for use at the Emory Cognitive Neurology Clinic, combining aspects from the Lawton and Brody Physical Self-Maintenance Scale, Lawton and Brody Instrumental Activities of Daily Living Scale, and an additional question about driving, this is the first study to examine harmonization techniques for FAQ and MAIADL Scale. Results will be useful to Emory clinicians measuring trajectory of functional impairment over time.

When we plotted the total FAQ scores versus the total MAIADL Scale scores, we noticed a nonlinear pattern and therefore tested for the significance of higher order terms. Through model selection, we chose the

polynomial regression with a linear and quadratic term for MAIADL Scale score and a linear term for date as the line of best fit. Because some individuals had ten years between the administration of the FAQ and MAIADL Scale instruments, and some individuals had only a few days, we wanted to control for this time difference in our model. We then applied this model to harmonize test scores for each individual, so we can measure trajectory of functional impairment over time. Therefore, it may not be realistic to harmonize scores that are far apart in time because later scores may be showing functional decline. We may be able to improve model accuracy by restricting the time between test administration for these two instruments.

After conducting item response theory (IRT), the discrimination estimates ranged from 3.42 to 18.21 for FAQ (a five fold increase) and from 2.66 to 7.23 for MAIADL Scale (an almost three fold increase). Because there is so much variety in item scores, we can be confident in saying IRT is necessary for harmonizing these two scales. The MAIADL Scale includes items which measure both ADLs (items from the Lawton and Brody Physical Self Maintenance Scale) and IADLs (items from the Lawton and Brody Instrumental Activities of Daily Living Scale). This further justifies our use of IRT because ADL items should not contribute to the overall score as much as IADLs given that the FAQ only measures IADLs. For these scores to be properly harmonized, they need to be measuring the same latent trait.

The items with the highest weights for FAQ were taxes (18.21), bills (17.36), and shopping (8.17), and the items with the highest weights for MAIADL Scale were shopping (7.23), food preparation (6.82), and ability to handle finances (5.79). There is overlap in the top three categories for the two instruments, which further confirms IRT is providing good harmonization. Intuitively it makes sense that these items differentiate the most between individuals of high functional ability versus low functional ability as they are difficult and important tasks for daily living.

We plotted the factor scores, representing functional ability, as well as the weighted scores using the discrimination estimates from IRT. The line of best fit for the factor scores included only the linear term for MAIADL Scale whereas the line of best fit for the weighted scores included both the linear and quadratic terms for MAIADL Scale. For factor scores, the concordance correlation coefficient (CCC) increases after eliminating people with more than five years between test administration dates and again after eliminating people with more than three years between test administration dates, implying harmonization is best when limiting the time between test dates. However, we do not see this same amount of change for the overall weighted scores. Therefore, there needs to be more work done to determine the optimal range between test dates before harmonization.

When comparing our different harmonization methods, we found that agreement and model accuracy was improved by using the weighted scores versus the original scores for FAQ and MAIADL Scale. Furthermore, CCC was highest and RMSE was lowest for factor scores (CCC=0.8585; RMSE=9.69) compared to both original scores (CCC=0.7320; RMSE=17.51) and weighted scores (CCC=0.7589; RMSE=17.37). This means there was greater agreement between and better model prediction using the estimated latent functional abilities based on FAQ and MAIADL Scale scores using IRT directly. Although factor scores yielded the best harmonization results, these estimates are difficult for clinicians to calculate and interpret. Therefore, weighted scores would be most practical to use in the clinic for comparing FAQ and MAIADL Scale scores.

Our results from performing IRT on the NACC Uniform Data Set (UDS) shows less variety in discrimination estimates for FAQ compared to using the NeuCog data set. The estimates range from 3.72 to 7.46, which is only a two fold increase. There are demographic and clinical differences in the two data sets, therefore in the future we would like to further explore these differences by conducting differential item functioning (DIF). DIF would allow us to capture the extent to which an item measures a latent trait differently for members of different subgroups. This could prove beneficial in better distinguishing how items contribute to overall functional ability. Additionally, to our knowledge this is the first study using IRT to test the efficacy of FAQ. Given that 15-20% of adults over 65 years old suffer from mild cognitive impairment which is diagnosed with the help of FAQ scores, this work has broad implications for clinical care of aging adults.

Lastly, we used survival to clinically evaluate our new weighted FAQ score because there is not gold standard method for how to do this. Although we found that FAQ is a significant predictor of time to death and time to dementia, using the weighted FAQ scores did not improve risk differentiation for either outcome. In the future, it could be beneficial to repeat our survival analyses using the factor scores, which directly measure functional ability. Additionally, we may need a better way to evaluate the impacts of our new weighted score as we saw from the IRT results that items do contribute differently to functional ability and should be considered when diagnosing patients.

# 5 Conclusion

In conclusion, individual test items contribute differently to measuring functional ability. Therefore, we should consider item scores as opposed to overall test scores when diagnosing individuals with cognitive

impairment. Factor scores had the best agreement between FAQ and MAIADL Scale compared to the original scores and the weighted scores. Additionally, the linear model using factor scores performed best for predicting functional ability measured by FAQ from functional ability measured by MAIADL Scale. While our new weighted FAQ scale was a significant predictor of both time to death and time to dementia after controlling for demographic and clinical covariates, it did not differentiate between high and low risk individuals better than the original FAQ scores in the NACC UDS. Our study results demonstrate how item response theory can be beneficial when measuring latent traits such as functional ability and we believe these methods can be applied in the future to better diagnose individuals with cognitive impairment.

# References

1. Alzheimer's Association (n.d.). *Mild Cognitive Impairment (MCI).* https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment?utm_source=google&utm_medium=paidsearch&utm_campaign=google_grants&utm_content=types_of_dementia&gclid=CjwKCi_D_BRApEiwASslbJ1Czc3TNv9JsLyK-FprmeUaFcLdytXjyN9rwlUoUZc89nE7mnlHjHBoCtFMQAvD_BwE

2. Alzheimer's Association (n.d). *What is Dementia?.* https://www.alz.org/alzheimers-dementia/what-is-dementia

3. Hall, J.R., Vo, H.T., Johnson, L.A., Barber, R.C., & O'Bryant, S.E. (2011). The link between cognitive measures and ADLs and IADL functioning in mild Alzheimer's: What has gender got to do with it? *International Journal of Alzheimer's Disease, 2011*, 2011:276734. https://doi.org/10.4061/2011/276734

4. Lawton, M.P. & Brody, E.M. (1969). Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living. *Gerontologist, 9*(3), 179-186.

5. Pfeffer, R.I., Kurosaki, T.T., Harrah, C.H., Chance, J.M., Filos, S. (1982). Measurement of Functional Activities in Older Adults in the Community. *Journal of Gerontology, 37*(3), 323-329.

6. Ward, G., Jagger, C., & Harper, W. (1998). A review of instrumental ADL assessments for use with elderly people. *Reviews in Clinical Gerontology, 8*(1), 65-71. https://doi.org/10.1017/S0959259898008089.

7. Malek-Ahmadi, M., Chen, K., Davis, K., Belden, et. al (2015). Sensitivity to change and prediction of global change for the Alzheimer's Questionnaire. *Alzheimer's Resaerch & Therapy, 7*(1). https://doi.org/10.1186/s13195-014-0092-z

8. McCutcheon, A.L.(1987). *Latent class analysis.* Sage Publications, Inc.

9. Bartholomew, D. J. Stell, F., Moustake, I., Galbraith, J.I. (2008). *Analysis of Multivariate Social Science Data, $2^{nd}$ Edition.* Chapman and Hall/CRC.

10. Crane, P. K., Narasimhalu, K., Gibbons, L. E., et al (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology, 61*(10), 1018-27.e9. https://doi.org/10.1016/j.jclinepi.2007.11.011.

11. Gross, A.L., Sherva, R., Mukherjee, S., et al (2014). Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology, 43*(0), 194-205. https://doi.org/

10.1159/000367970

12. Balsis, S., Unger, A. A., Benge, J. F., Geraci, L., Doody, R. S. (2012). Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: A comparison of item response theory-based scores and total scores. *Alzheimer's & Dementia, 8*(4), 288-294. https://doi.org/10.1016/j.jalz.2011.05.2409

13. Gross, A.L., Power, M.C., Albert, M.S., et al (2015). Application of latent variable methods to the study of cognitive decline when tests change over time. *Epidemiology, 26*(6), 878-887. https://doi.org/10.1097/EDE.0000000000000379

14. Chan, S.k, Gross, A.L., Pezzin, L.E., Brandt, J., Kasper, J.D. (2015). Harmonizing measures of cognitive performance across international surveys of aging using item response theory. *Journal of Aging and Health, 27*(8), 1392-1414. https://doi.org/10.1177/0898264315583054

15. Morris, J.C., Weintraub, S. Chui, H.C, et al (2006). The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease & Associated Disorders, 20*(4), 210-216. https://doi.org/10.1097/01.wad.0000213865.09806.92

16. Vittengl, J.R., White, C.N., McGovern, R.J., & Morton, B.J. (2006). Comparative validity of seven scoring systems for the instrumental activities of daily living scale in rural elders. *Aging and Mental Health, 10*(1), 40-7. https://doi.org/10.1080/13607860500307944

17. Kendall, M. (1938). A new measure of rank correlation. *Biometrika, 30*(1-2), 81-89. https://doi.org/10.1093/biomet/30.1-2.81

18. SAS [computer program]. Version 9.4. Cary, NC: SAS Institute Inc; 2014.

19. Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255-268. https://doi.org/10.2307/2532051

20. Hableton, R.K., Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of item response theory.* Sage Pubilcations: Newbury Park, CA

21. Kline, R.B. (2004). *Principals and practice of structural equation modeling.* The Guilford Press: New York, NY

22. Tucker, L. R., Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10. https://doi.org/10.1007/BF02291170

23. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (Psychometric Monograph No. 17) Richmond, VA: Richmond Society.

24. SAS Institute Inc. 2015. SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

25. R: A Language and Environment for Statistical Computing [computer program]. Version R 4.0.2. Vienna, Austria: R Foundation for Statistical Computing; 2020.

26. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B., 34*(2), 187-220.

# Appendix

**Supplemental Table 1. Emory Cognitive Neurology MAIADL Scale Adapted from Lawton & Brody, 1969[4]**

| Item | Score |
|---|---|
| 1A. Physical Self-Maintenance Scale | |
| A. Toilet | |
| Cares for self at toilet completely, no incontinence | 0 |
| Needs to be reminded, or needs help in cleaning self, or has rare (weekly at most) accidents. | -1 |
| Soiling or wetting while asleep more than once a week. | -2 |
| Soiling or wetting while awake more than once a week. | -3 |
| No control of bowels or bladder. | -4 |
| B. Feeding | |
| Eats without assistance. | 0 |
| Eats with minor assistance at meal times and/or with special preparation of food, or help in cleaning up after meals. | -1 |
| Feeds self with moderate assistance and is untidy. | -2 |
| Requires extensive assistance for all meals. | -3 |
| Does not feed self at all and resists efforts of others to feed him. | -4 |
| C. Dressing | |
| Dresses, undresses, and selects clothes from own wardrobe. | 0 |
| Dresses and undresses self, with minor assistance. | -1 |
| Needs moderate assistance in dressing or selection of clothes. | -2 |
| Needs major assistance in dressing, but cooperates with efforts of others to help. | -3 |
| Completely unable to dress self and resists efforts of others to help. | -4 |
| D. Grooming (neatness, hair, nails, hands, face, clothing) | |
| Always neatly dressed, well-groomed, without assistance. | 0 |
| Grooms self adequately with occasional minor assistance, e.g., shaving. | -1 |
| Needs moderate and regular assistance or supervision in grooming. | -2 |
| Needs total grooming care, but can remain well-groomed after help from others. | -3 |
| Actively negates all efforts of others to maintain grooming. | -4 |
| E. Physical Ambulation | |
| Goes about grounds or city. | 0 |
| Ambulates within residence or about one block distant. | -1 |
| Ambulates with assistance of (check one) a) another person, b) railing, c) cane, d) walker, e) wheel chair.<br>1__Gets in and out without help.<br>2__Needs help in getting in and out | -2 |
| Sits unsupported in chair or wheelchair, but cannot propel self without help. | -3 |
| Bedridden more than half the time. | -4 |
| F. Bathing | |
| Bathes self (tub, shower, sponge bath) without help. | 0 |
| Bathes self with help getting in and out of tub. | -1 |
| washes face and hands only, but cannot bathe rest of body. | -2 |
| Does not wash self but is cooperative with those who bathe him. | -3 |

| | |
|---|---|
| Does not try to wash self and resists efforts to keep him clean. | -4 |
| 1B. Instrumental Activities of Daily Living Scale | |
| A. Ability to Use Telephone | |
| Operates telephone on own initiative–looks up and dials numbers, etc. | 0 |
| Dials a few well-known numbers. | -1 |
| Answers telephone but does not dial. | -2 |
| Does not use telephone at all. | -3 |
| B. Shopping | |
| Takes care of all shopping needs independently. | 0 |
| Shops independently for small purchases. | -1 |
| Needs to be accompanied on any shopping trip. | -2 |
| Completely unable to shop. | -3 |
| C. Food Preparation | |
| Plans, prepares, and serves adequate meals independently. | 0 |
| Prepares adequate meals if supplied with ingredients. | -1 |
| Heats and serves prepared meals, or prepares meals but does not maintain adequate diet. | -2 |
| Needs to have meals prepared and served. | -3 |
| D. Housekeeping | |
| Maintains house alone or with occasional assistance (e.g., "heavy work-domestic help"). | 0 |
| Performs light daily tasks such as dishwashing, bedmaking. | -1 |
| Performs light daily tasks but cannot maintain acceptable level of cleanliness. | -2 |
| Needs hlep with all home maintenance tasks. | -3 |
| Does not participate in any housekeeping tasks. | -4 |
| E. Laundry | |
| Does personal laundry completely. | 0 |
| Launders small items - rinses socks, stockings, etc. | -1 |
| All laundry must be done by others. | -2 |
| F. Mode of Transportation | |
| Travels independently on pubic transportation or drives own car. | 0 |
| Arranges own travel via taxi, but does not otherwise use pubic transportation. | -1 |
| Travels on public transportation when assisted or accompanied by another. | -2 |
| Travel limited to taxi or automobile with assistance of another. | -3 |
| Does not travel at all. | -4 |
| G. Responsibility for own Medications | |
| Is responsible for taking edication in correct dosages at correct time. | 0 |
| Takes responsibility if medication is prepared in advance in separate dosages. | -1 |
| Is not capable of dispensing own medication. | -2 |
| H. Ability to Handle Finances | |
| Manages financial matters independently (budgets, writes checks, pays rent, bills, goes to bank), collects and keeps track of income. | 0 |
| Manages day-to-day purchases, but needs help with banking, major purchases, etc. | -1 |
| Incapable of handling money. | -2 |
| I. Driving | |
| Drives alone safely. | 0 |
| Drives alone but has had one or more recent accidents. | -1 |

| | |
|---|---|
| Drives alone but has gotten lost. | -2 |
| Drives only with someone else in the car. | -3 |
| Never drives. | -4 |

**Supplemental Table 2. NACC Functional Assessment Scale**[5]

| In the past four weeks, did the subject have difficulty or need help with: | Not Applicable (e.g., never did) | Normal | Has difficulty, but does by self | Requires assistance | Dependent | Unknown |
|---|---|---|---|---|---|---|
| 1. Writing Checks, paying bills, or balancing a check book | 8 | 0 | 1 | 2 | 3 | 9 |
| 2. Assembling tax records, business affairs, or other papers | 8 | 0 | 1 | 2 | 3 | 9 |
| 3. Shopping alone for clothes, household necessities, or groceries | 8 | 0 | 1 | 2 | 3 | 9 |
| 4. Playing a game of skill such as bridge or chess, working on a hobby | 8 | 0 | 1 | 2 | 3 | 9 |
| 5. Heating water, making a cup of coffee, turning off the stove | 8 | 0 | 1 | 2 | 3 | 9 |
| 6. Preparing a balanced meal | 8 | 0 | 1 | 2 | 3 | 9 |
| 7. Keeping track of current events | 8 | 0 | 1 | 2 | 3 | 9 |
| 8. Paying attention to and understanding a TV program, book, or magazine | 8 | 0 | 1 | 2 | 3 | 9 |
| 9. Remembering appointments, family occasions, holidays, medications | 8 | 0 | 1 | 2 | 3 | 9 |
| 10. Traveling out of the neighborhood, driving, or arranging to take public transportation | 8 | 0 | 1 | 2 | 3 | 9 |