Psychological Test and Assessment Modeling, 2026.68:3-31
DOI: https://doi.org/10.2440/001-0024

Development of a Cognitive Screen Item Bank: Assessment of Individuals Across the Cognitive Spectrum, Including Advanced Dementia

Jeanne A. Teresi[1,2,3,4], Katja Ocepek-Welikson[2], Marjorie Kleinman[4], Mildred Ramirez[2,3], Joseph P. Eimicke[2], Stephanie Silver[2], Jose Luchsinger[2], Albert Siu[5], Dan Mroczek[6], David Cella[6]

1 Columbia University Stroud Center, New York, NY, USA
2 Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA
3 Department of Geriatrics and Palliative Medicine, Weill Cornell Medical Center, New York, NY, USA
4 New York State Psychiatric Institute, New York, NY, USA
5 Department of Geriatrics and Palliative Medicine, General Internal Medicine, Health Evidence and Policy, Mount Sinai Medical Center, New York, NY USA
6 Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Abstract:
An item bank of cognitive items targeting the spectrum of cognitive function, with a focus on advanced dementia was developed with data from residents in long term services and supports (LTSS) settings. Latent variable models, including item response theory and factor analyses were applied to fifty items from multiple cognitive screening measures assessing memory, orientation, naming, attention, calculation, and following directions. Because the intent was to use the item bank to measure overall cognition, subdomains were not modeled. Factor analyses tested for essential unidimensionality, followed by application of an item response theory-based graded response model, including differential item functioning to examine measurement equivalence. The total sample size was 6,921, with 1,874 (27.1 %) males and 5,040 (72.9 %) females. The average age was 82.5 (*SD* = 11.0). There were 1,089 (16.2 %) Black, 498 (7.4 %) Hispanic, and 5,136 (76.4 %) White respondents. The average number of years of education was 9.1 (*SD* = 5.5). Estimates of reliability across age, education, race, and sex subgroups were high with most > 0.96. Cognitive function of the participants in the study varied from no to minimal

Correspondence:

Jeanne A. Teresi, Columbia University Stroud Center at New York State Psychiatric Institute; teresimeas@aol.com; jat61@cumc.columbia.edu

impairment (552 or 8.0 %); mild (2,000 or 28.9 %); moderate (2,686 or 38.8 %); severe (1,328 or 19.2 %) to very severe impairment (355 or 5.1 %). Items with low information and substantial DIF were recommended for removal. The final recommended item bank includes 38 items. Future work with this item bank may include development of computerized adaptive tests and short-forms.

Keywords: item bank, cognition, dementia, item response theory, factor analysis

Several health-related item banks that also contain measures of cognition exist; however, most of these efforts have been focused on applied cognition and have not included populations of advanced age with more advanced cognitive impairment or those living in long term services and supports (LTSS) settings. Examples include Neuro-QOL, (Gershon, Lai, Bode, Choi, Moy, Bleck, Miller, Peterman, & Cella, 2012), the Patient-Reported Outcomes Measurement Information System (PROMIS; Cella, Riley, Stone, Rothrock, Reeve, Yount, Amtmann, Bode, Buysse, Choi, Cook, DeVellis, DeWalt, Fries, Gershon, Hahn, Pilkonis, Revicki, Rose, Weinfurt, & Hays, 2010; Cella, Yount, Rothrock, Gershon, Cook, Reeve, Ader, Fries, Bruce, & Rose, 2007; Reeve, Hays, Bjorner, Cook, Crane, Teresi, Thissen, Revicki, Weiss, Hambleton, Liu, Gershon, Reise, Lai, & Cella, 2007), the Activity Measure for Post-Acute Care (AMPAC; Haley, Coster, Andres, Ludlow, Ni, Bond, Sinclair, & Jette, 2004) and more recently, the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT) applied cognition item bank. The latter bank, focused on applied cognition general concerns, including functional deficits associated with cognitive complaints was described and evaluated for DIF (Cheville, Wang, Yost, Teresi, Ramirez, Ocepek-Welikson, Ni, Marfeo, Keeney, Basford, & Weiss, 2021; Teresi, Ocepek-Welikson, Ramirez, Kleinman, Wang, Weiss, & Cheville, 2022). The NIH Toolbox (Fox, Zhang, Amagai, Bassard, Dworak, Han, Kassanits, Miller, Nowinski, Giella, Stoeger, Swantek, Hook, & Gershon, 2022; Gershon, Wagster, Hendrie, Fox, Cook, & Nowinski, 2013) contains a cognitive test battery, including attention, working memory, processing speed and crystallized cognitive abilities; however, this bank has not included many individuals of advanced age or from LTSS settings.

Aims

The goal of this project was to develop an item bank of cognitive screening items for use across all stages of dementia, including advanced dementia. Latent variable item response theory and factor analyses models were used. The aims of these analyses were to: 1) compile a pool of cognitive screening items from individuals in LTSS; 2) examine the dimensionality and reliability of the pool; 3) perform differential item functioning (DIF); and 4) based on the analyses of DIF and examination of the

information provided by the items, select a final set of items that may comprise an item bank for potential use in computerized adaptive testing (CAT) and construction of short-forms. This set of analyses may serve as a reference for professionals engaged in the development of CATs and short-forms for use with LTSS populations. Ultimately, the item bank would be available electronically in the medical record assessment protocol for use by clinicians. An item bank containing screening items could produce more efficient screening in applied settings for which economies of time and effort are prioritized. Such settings could include clinics, hospitals, home health care as well as institutional LTSS.

Differential Item Functioning

An important step in selecting items for item banks is to examine measurement equivalence, to identify a form of item bias. Socio-demographic characteristics including education, age, and racial and ethnic background can result in differential performance on cognitive items such that group differences in item response do not reflect true group difference in cognitive ability. Item banks are constructed after removing items with high impact DIF or providing separate calibrations for certain groups. An alternative is to flag selected items as "enemy" items that are not administered. Given the extensive previous research examining some of the items used commonly, when items were identified here with DIF congruent with that identified in earlier studies, the practice was to remove the items from the bank. Research conducted over the past three decades has demonstrated the necessity of examining measurement equivalence in cognitive measures. Previous analyses have focused on DIF in cognitive screening measures (Crane, van Belle, & Larson, 2004; Stump, Monahan, & McHorney, 2005) and other neuropsychological tests with respect to race/ethnicity and education (Teresi, Holmes, Ramirez, Gurland, & Lantigua, 2001). For example, screening measures have evidenced DIF by education (Escobar, Burnam, Karno, Forsythe, Landsverk, & Golding, 1986; Jones & Gallo, 2002; Stump et al., 2005; Teresi, Golden, Cross, Gurland, Kleinman, & Wilder, 1995; Valle, Hough, Kolody, Cook-Gait, Velazquez, & Jimenez, 1991). Those with less education at the same level of cognition than those with higher education were more likely to answer some items incorrectly. Groups with lower education as contrasted with those with higher education have evidenced greater difficulty with items requiring the performance of math (e.g., serial subtraction) and language-based (e.g., spelling backwards, write a sentence) tasks. DIF in cognitive test items has also been observed for sex. Jones and Gallo (2002) detected sex-DIF in a cognitive screening measure, concluding that a large proportion of sex differences in cognitive impairment were explained by item bias. Women showed more difficulty with serial subtractions, and men on language-related tasks. Similarly, Dubbelman and colleagues (2020), examining functional impairment among memory clinic patients in eight countries, documented sex DIF on items related to tasks stereotypically associated with sex-norms.

Education DIF has been observed in neuropsychological measures (Gibbons, Crane, Mehta, Pedraza, Tang, Manly, Narasimhalu, Teresi, Jones, & Mungas, 2011; Mungas, Widaman, Reed, & Tomaszewski Farias, 2011). DIF in subjective cognitive assessment and applied cognition measures has also been observed (Teresi, Kleinman, Ocepek-Welikson, 2000). Fieo, Ocepek-Welikson, Kleinman, Eimicke, Crane, Cella, and Teresi (2016) examined DIF in the Patient Reported Outcomes Measurement Information System (PROMIS®; Cella et al., 2007; Reeve et al., 2007); in that study little DIF of high magnitude was observed. One item showed DIF by several methods, and with a higher magnitude for race/ethnicity: "It has seemed like my brain was not working as well as usual".

DIF has been documented in measures used frequently to assess cognition (Crane et al., 2004; Fieo et al., 2016; Filshtein, Chan, Mungas, Whitmer, Fletcher, DeCarli, & Farias, 2020; Stump et al., 2005; Teresi et al., 2001). For instance, items from the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), a widely used cognitive screening measure, have evidenced DIF with respect to race and/or ethnicity, language, and education level (see Escobar et al., 1986; Hohl, Grundman, Salmon, Thomas, & Thal, 1999; Jones & Gallo, 2002; Marshall, Mungas, Weldon, Reed, & Haan, 1997; Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006; Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Teresi et al., 1995; Valle et al., 1991). An in-depth discussion about putative causes of DIF on specific MMSE and other cognitive screening items can be found in Ramirez, Teresi, Holmes, Gurland, and Lantigua (2006). Modification based on DIF analyses i.e., item removal or separate calibrations, may be pertinent contingent upon the use and application of the specific measure (Teresi, Ramirez, Jones, Choi, & Crane, 2012).

Method

Description of the Item Pool

Data for this item pool was obtained from residents in LTSS settings, including home care, day care, independent living, assisted living and nursing homes. Fifty items with the most complete responses from all studies were selected for the pool.

Items were taken from multiple cognitive screening measures and were posited to be essentially unidimensional. The potential subdomains represented included memory (e.g., remembering age, month, year of birth); orientation (knowing the name of the facility, the city, town and state, the month and year); naming (naming a pencil, watch); attention (spelling world backwards, counting backwards from 20 to 1, counting backwards from 100 by 7); calculation (making change); following directions (fold a paper, following simple commands). Because the intent was to use the item bank to measure overall cognition including in advanced dementia and among individuals in LTSS settings, subdomains were not modeled and as shown below, strong support for unidimensionality of the item set was observed.

Cognitive Assessment Scales

Scales from which items were obtained are described briefly. The Comprehensive Assessment and Referral Evaluation (CARE) Mental Status Questionnaire (CMSQ) was developed cross-nationally for use in community settings (Copeland, Kelleher, Kellet, Gourlay, Gurland, Fleiss, & Sharpe, 1976; Golden, Teresi, & Gurland, 1983; Golden, Teresi, & Gurland, 1984; Gurland, Fleiss, Goldberg, Sharpe, Copeland, Kelleher, & Kellet, 1976; Gurland, Golden, Teresi, & Challop, 1984; Gurland, Kuriansky, Sharpe, Simon, Stiller, & Birkett, 1977-1978; Gurland, & Wilder, 1984; Teresi, Golden, & Gurland, 1984; Teresi, Golden, Gurland, Wilder, & Bennett, 1984. The INCARE (institutional version of the CARE) was developed from a series of cross-national institutional studies conducted from 1977-1979 (Mann, Wood, Cross, Gurland, Schreiber, & Haefner, 1984). The INCARE Cognitive Screening Measure determines the ability of the individual to respond to increasingly complex questions. Included in the screen are assessments of (a) arousal, (b) level of alertness, simple commands, (d) cognitive functioning, (orientation, memory, calculation / attention), (e) range of motion and ambulation, and (f) affect. The psychometric properties are good to excellent across many community and institutional U.S. and cross-national samples.

The measure contains items from five cognitive screening measures: 1) The Blessed Memory-Information-Concentration Test (Blessed, Tomlinson, & Roth, 1968) contains 26 items measuring orientation, memory, and calculation/attention. Typical items are: "cannot spell own name," "does not recall dates for World War I," "does not know correct hour of the day." 2) The ten-item Kahn-Goldfarb Mental Status Questionnaire (MSQ; Kahn, Goldfarb, Pollack, & Gerber, 1960) contains items such as: "does not know own age," "does not know year of birth," "does not know today's date." 3) The Standardized Mini-Mental State Examination (SMMSE; Folstein et al., 1975; Molloy, Alemayehu, & Roberts, 1991) contains 20 items measuring attention, calculation, recall, and language. The SMMSE tests also the ability to follow verbal and written commands. Memory is measured by asking the respondent to recall a list of three objects memorized earlier. Typical orientation items are: "cannot name city s/he is in," "does not know the current year." Attention is measured by asking the respondent to spell the word "world" backwards or to calculate "serial 7's" (serial 7s is not in the SMMSE but was included from earlier studies in this data set), and to follow the commands for folding a piece of paper. 4) The Short Portable Mental Status Questionnaire (SPMSQ; Pfeiffer, 1975) contains ten items tapping cognitive ability. Typical items are "does not know place of birth," "does not know name of facility," "does not know day of the week." 5) The Care Diagnostic Scale (CAREDIAG) has been studied using several advanced psychometric models, including analyses of its relationship to dementia diagnosis (Gurland, Wilder, Cross, Teresi, & Barrett, 1992; Teresi, Kleinman, Ocepek-Welikson, Ramirez, Gurland, Lantigua, & Holmes, 2000). This scale has been found to be more culturally fair than others (Ramirez et al., 2006; Teresi, 2007). Nine items in the bank were similar in content to the Montreal

Cognitive Assessment (MoCA; Nasreddine, Phillips, Bedirian, Charbonneau, Whitehead, Collin, Cummings, & Chertkow, 2005).

Item Responses. As described earlier, fifty unique items with the most complete responses were used. Nevertheless, items evidenced missing data for a variety of reasons, including non-response because the subjects were not continuously alert. A special 'missing' response category (7 – 'Response not ratable') was used and recoded into an additional worst valid response (2 in the case of binary items). Another non-response was a refusal to answer a question, the lowest response option, which in keeping with clinical practice in cognitive assessment was recoded into an incorrect response. No other missing responses were recoded or prorated. Items were coded as 0 for "correct", 1 for "incorrect" and 2 for "no response or response not rateable". The latter category reflects a higher level of cognitive incapacity and includes answers that are off the point, or not given due to end-stage impairment affecting communication ability. There were also nine items with counts of errors that have more than 3 and up to 7 categories.

Psychometric Approach

Unidimensionality and Local Independence

Although cognitive subdomains were represented by the item pool, the goal was to determine if the resulting item bank could be used to select items across subdomains for assessment. Thus, model assumptions were examined with respect to the entire item pool. Unidimensionality is an assumption underlying item response theory models used to develop CATs, and when performing the DIF analyses. Essential unidimensionality was examined through a merged exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Asparouhov & Muthén, 2009) performed by fitting a unidimensional model with polychoric correlations using Mplus (Muthén & Muthén, 2011). The comparative fit index (CFI) and root mean square error of approximation (RMSEA) were evaluated as indicators of fit (Bentler, 1990). The explained common variance (ECV), estimated as the percent of observed variance explained (Reise, Moore & Haviland, 2010; Rodriguez, Reise, & Haviland, 2016) informed whether the observed variance covariance matrix was close to unidimensionality (Sijtsma, 2009).

Local independence, another IRT assumption was assessed with the generalized, standardized local dependency chi-square statistics (Chen & Thissen, 1997) provided in IRTPRO (Cai, Thissen, & du Toit, 2011).

Model for IRT

Because the items were ordinal, the analyses were conducted using the graded response model (Samejima, 1969). Ordered responses, $x = k$ and $k = 1,2,...m$ are assumed, in which a_i is the discrimination (slope) for item i and b_{ik} the difficulty parameters for response category k:

$$P(x = k) = P^*(k) - P^*(k+1) = 1 / 1 + \exp[-a_i(\theta - b_{ik})] - 1 / 1 + \exp[-a_i(\theta - b_{ik+1})].$$

$P^*(k)$ is the item characteristic curve (ICC) describing the probability that a response is in category k or higher, for each value of θ (see also Orlando-Edelen et al., 2006; Thissen, 1991).

There are $k-1$ boundary response functions describing the cumulative probability of responding in category k or higher. The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, e.g., cognition (denoted θ), measured by the item set is characterized by location (severity) parameter(s) (denoted b) and a discrimination parameter (denoted a), proportional to the slope of the curve.

DIF Assessment Using IRT

One method for examining measurement equivalence is DIF analyses, which in this set of analyses was focused on differences in item performance after conditioning on a cognitive trait. DIF occurs when the probability of item category endorsement differs across comparison groups when holding the trait constant. DIF is demonstrated by parameters and ICCs that are different across the subgroups examined. If tests of the equivalence of the a parameters (indicative of non-uniform DIF) are not significant, tests of group differences in the b parameters (indicating uniform DIF) are performed, constraining the a parameters to be equal.

The item response theory likelihood ratio (IRTLR) method tests a series of IRT models established by fixing and freeing parameters. The method for DIF detection was a comparison of parameters from nested models using a variant of the Wald test based on Lord's chi-square (Cai et al., 2011; Langer, 2008; Lord, 1980; Teresi, Kleinman, & Ocepek-Welikson, 2000; Woods, Cai, & Wang, 2013). The Wald test for DIF follows the model proposed by Lord (1980) in which vectors of IRT item parameters are compared; the model was extended for polytomous data by Cohen, Kim, and Baker (1993). The Wald test is asymptotically equivalent to the likelihood ratio test (Thissen, 1991). IRTPRO (Cai et al., 2011) software was used for the DIF analyses. Age, sex, education, and race/ethnicity DIF were examined.

Anchor Items. The comparison groups were linked on the cognition construct, and the mean and variance estimated for the target groups studied (while setting the reference group mean to 0 and variance to 1). After the initial determination of items with DIF, the anchor set was studied iteratively (Orlando-Edelen et al., 2006; Wang, Shih, &

Sun, 2012; Woods, 2009). Items showing DIF were excluded from the anchor set at each iteration until no items showed DIF, and this set was used for final determination of DIF. A model was constructed with all parameters constrained to be equal across comparison groups for the anchor items, and item parameters for all studied items freed to be estimated distinctly. An overall simultaneous joint test of differences in the *a* or *b* parameters was performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF was detected when the *b* parameters differed and non-uniform DIF when the *a* parameters differed. Orthogonal contrasts were used. The final *p* value was adjusted for examination of number of items in the analysis, i.e., 50 items (final significant $p \leq 0.001$; Bonferroni, 1936).

Evaluation of DIF Magnitude and Effect Sizes. Expected item scores were examined as measures of magnitude. An expected item score is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible categories for the item. The method used for quantification of the difference in the average expected item scores was the non-compensatory DIF index (NCDIF; Raju, van der Linden, & Fleer, 1995). Cutoff values established based on simulations (Fleer, 1993; Flowers, Oshima, & Raju, 1999) were used in the estimation of the magnitude of item-level DIF. The square root of NCDIF provides an effect size in terms of the original metric. For example, the cutoff values were 0.0540 for polytomous items with four response options, 0.0240 for three and 0.0060 for two response categories (Raju, 1999). Thus, for a polytomous item with four response categories, the recommended cutoff corresponds to an average absolute difference of 0.232 (almost one quarter of a point) on a four-point scale (see Raju, 1999; Meade, Lautenschlager, & Johnson, 2007). An additional effect size measure (T1) proposed by Wainer (1993) and extended for polytomous data by Kim, Cohen, Alagoz, and Kim (2007) was also examined with a recommended cutoff of 0.10 (Wainer, 1993). T(1) is the sum of the differences both positive and negative, and divided by N. The theta and item parameter estimates for the two groups were equated using iterative purification of equating constants (Seybert & Stark, 2012) and placed on a common metric. See Kleinman and Teresi (2016) for a detailed description of the procedures, and Orlando-Edelen, Stucky, and Chandra (2015), Chalmers, Counsell, and Flora (2016), and Chalmers (2018) for extensions of this approach.

Evaluation of DIF Impact. Aggregate-level impact was evaluated, examining expected scale score functions. Expected item scores were summed to produce an expected scale score which provides evidence regarding the effect of DIF on the total score. Group differences in these scale response functions provide overall aggregated measures of DIF impact. (See Kleinman & Teresi, 2016; Teresi, Wang, Kleinman, Jones, & Weiss, 2021). Final DIF designations were based on three criteria: (1) Wald test showing a significant DIF in the *'a'* and/or *'b'* parameters after an adjustment for multiple tests; (2) NCDIF test above the recommended threshold; and (3) the T(1) test if it exceeds 0.1.

Evaluation of Reliability and Information. Reliability was evaluated by decomposing the scale score into the sum of the item scores, and the contribution of the common term or communality. McDonald's (McDonald, 1999) Omega Total (ω_t), a reliability estimate that is based on the proportion of total common variance explained, was also calculated. Both Cronbach's alpha (Cronbach, 1951) and ordinal alpha based on polychoric correlations (Zumbo, Gadermann, & Zeisser, 2007) were calculated. Additionally, IRT-based reliability measures were examined at selected points along the underlying latent continuum. Finally, the item and test information functions were calculated and graphed.

Missing Responses. As described earlier, consistent with clinical applications of screening scale items, two types of missing data were coded as valid responses: a) a non-response on an item because a patient was not continuously alert through the interview or b) a patient refused to answer an item. Dimensionality analyses were conducted with the sample which included the remainder of missing data. The assumption for missing data was missing at random; therefore, pairwise estimation of the covariance matrix using weighted least square estimation was used. The approach is explained in the Mplus v.6 manual (Mplus, Chapter 14, Special Modeling Issues, p.435). Sensitivity analyses were performed using listwise deletion to provide confirmation of the results of factor analyses described below.

IRTPRO implements the method of Maximum Likelihood (ML) and expectation maximization (EM) item parameter estimation (Bock-Aitkin EM is the default; IRTPRO v.5 Guide, p.5; IRTPRO-Estimation-Methods.pdf, p.3) while the score was estimated by the expected a posteriori (EAP) estimator (IRTPRO v.5 Guide, p. 209). Therefore, DIF analyses were also conducted with the total sample including missing data.

Correlations Among the Original Scales and Theta Estimates

Pearson correlations of the 50-item summary scores and theta estimates of cognitive functioning with the five original measure sum-scores, BLESSED, SMMSE, CMSQ, MSQ, and SPMSQ were calculated.

Results

Sample

The total sample size was 6,921 which included item-level missing responses. The sample sizes were reduced with listwise deletion (see Table 1 for details).

Levels of cognitive function were based on theta estimates associated with the 50-item set and varied from no to minimal impairment (at $\theta < -1.5$; n = 552 or 8.0%); mild ($-1.5 < \theta < -0.5$; n = 2,000 or 28.9 %); moderate ($-0.5 < \theta < 0.5$; 2,582 or 37.3 %); severe ($0.5 < \theta < 1.5$; 1,432 or 20.7 %) to very severe impairment ($\theta \geq 1.5$; 355 or 5.1 %).

In terms of the studies and settings represented, there were a total of 18 studies (14 in nursing homes, n = 5,395 or 78.0 %; 2 in Assisted Living facilities, n = 778 or 11.2 %; 2 community samples (home health care and day care), n = 748 or 10.8 %). Data from the different samples were collected between January 1992 and February 2020.

Tests of Model Assumptions

Unidimensionality There was strong support for the unidimensionality of the item set for both the pairwise and listwise deletion of data approaches. The principal components analyses showed that the ratio of component 1 to 2 was very large (11.6 and 11.5) and the first component explained 56 % of the variance for both approaches to modelling missing data.

An examination of the model fit statistics was used as an additional test of dimensionality. Model fit indices supported the essential unidimensionality of the item pool set. Results of the CFA (Mplus) analysis with pairwise deletion of missing data for the unidimensional and two factor CFA were: CFI = 0.930 and 0.962; RMSEA = 0.068 and 0.051. Model fit statistics for the listwise deletion of missing data for the unidimensional and two factor CFA were: CFI = 0.929 and 0.964; RMSEA = 0.067 and 0.049. The two factor CFA did not improve the model fit appreciably for either approach to treatment of missing data (see Table 2). The ECV for the total sample was 83.6 (see Table 3).

Table 1.
Cognition Item Pool: Demographics

Sample characteristics	Total *n*	Percent	Listwise deletion of data	Percent
Total sample	6,921		3,504	
Age				
≤ 65	520	7.6	224	6.4
66 - 75	904	13.1	449	12.9
76 - 85	2,304	33.5	1,235	35.5
86 - 95	2,715	39.5	1,412	40.5
> 95	436	6.3	163	4.7
Total age	6,879	100.0	3,483	100.0
Average age (*SD*)	82.5 (11.0)		82.3 (11.8)	
Minimum age	20		20	
Maximum age	107		107	
Education				
No education	804	16.5	474	18.3
Grades 1 - 8	1,132	23.2	527	20.3
Grades 9 - 12	1,971	40.3	1,038	40.0
More than high school	978	20.0	558	21.5
Total education	4,885	100.0	2,597	100.0
Average number of years (*SD*)	9.1 (5.5)		9.2 (5.3)	
Minimum years	0		0	
Maximum years	22		22	
Race / ethnicity				
Black	1,089	16.2	469	13.7
Hispanic	498	7.4	186	5.4
White	5,136	76.4	2,773	80.9
Total race / ethnicity	6,723	100.0	3,428	100.0
Sex				
Male	1,874	27.1	923	26.3
Female	5,040	72.9	2,581	73.7
Total Sex	6,914	100.0	3,504	100.0

Table 2
Cognition Item Pool: Factor Loadings/Structure for One and Two Factor Solutions by Total Sample (Mplus v.6.1, Weighted Least Squares Mean and Variance (WLSMV) Estimation, OBLIMIN Rotation)

Item #	Item name	Item label	Pairwise deletion			Listwise (n = 3,504)		
			1 Fac-	2 Factors		1 Fac-	2 Factors struc-	
			λ (*SE*)	λ 1	λ 2	λ (*SE*)	λ 1	λ 2
1	BLESS1R	Spelling name	0.74	0.67	0.75	0.76	0.69	0.77
2	MSQ6R	Age	0.75	0.76	0.66	0.76	0.77	0.67
3	MSQ7R	Month of birth	0.75	0.71	0.73	0.75	0.72	0.73
4	MSQ8R	Year of birth	0.72	0.7	0.68	0.71	0.69	0.68
5	SPMSQ2R	Birthplace	0.67	0.64	0.65	0.64	0.59	0.64
6	MSQ1R	Where are we	0.92	0.93	0.62	0.92	0.94	0.66
7	S22R	What do people	0.92	0.93	0.60	0.91	0.93	0.64
8	S23R	What is the name	0.82	0.84	0.69	0.81	0.84	0.68
9	S23AR	What floor are we	0.79	0.79	0.71	0.79	0.78	0.72
10	MSQ2R	Street address of	0.75	0.78	0.62	0.75	0.78	0.63
11	BLESS8R	What city (town) is	0.69	0.72	0.58	0.70	0.73	0.6
12	SMMSE4R	What county are	0.71	0.73	0.63	0.70	0.72	0.63
13	S24CR	What state are we	0.75	0.74	0.7	0.77	0.74	0.74
14	BLESS7R	How long at this	0.66	0.69	0.56	0.67	0.7	0.56
15	MSQ4R	Current month	0.85	0.87	0.73	0.86	0.87	0.75
16	MSQ3R	Today's date	0.79	0.82	0.66	0.80	0.83	0.67
17	BLESS12R	Day of the week	0.82	0.84	0.7	0.81	0.84	0.69
18	MSQ5R	Current year	0.87	0.87	0.76	0.86	0.87	0.77
19	BLESS15R	Current season	0.79	0.78	0.74	0.79	0.78	0.73
20	BLESS14R	Hour of the day	0.78	0.77	0.73	0.78	0.76	0.73
21	BLESS13R	Part of the day	0.80	0.77	0.77	0.81	0.77	0.78
22	S57R	Test of touching	0.73	0.64	0.76	0.75	0.63	0.79
23	S60R	Recall interview-	0.68	0.71	0.57	0.68	0.72	0.57
24	CMSQ10R	Interviewer's	0.76	0.78	0.66	0.76	0.78	0.65
25	SMMSE7R	No ifs, ands, or	0.59	0.5	0.63	0.57	0.48	0.61
26	SMMSE8R	Name pencil	0.76	0.64	0.79	0.82	0.71	0.83
27	SMMSE9R	Name wristwatch	0.81	0.71	0.83	0.85	0.75	0.86

Item #	Item name	Item label	Pairwise deletion			Listwise (n = 3,504)		
			1 Fac-	2 Factors		1 Fac-	2 Factors struc-	
			λ (*SE*)	λ 1	λ 2	λ (*SE*)	λ 1	λ 2
28	SMMSE10R	Spell world back-	0.76	0.66	0.79	0.74	0.63	0.78
29	SMMSE11R	Close your eyes	0.75	0.66	0.76	0.74	0.65	0.76
30	SMMSE12R	Folding paper	0.64	0.56	0.67	0.61	0.52	0.64
31	SMMSE13R	Write sentence	0.77	0.65	0.8	0.76	0.64	0.8
32	SMMSE14R	Drawing pentagon	0.60	0.5	0.64	0.55	0.45	0.6
33	BLESS16R	Father's first	0.64	0.54	0.67	0.64	0.55	0.67
34	BLSS16AR	How much	0.61	0.55	0.62	0.60	0.54	0.61
35	BLESS17R	Name of a school	0.69	0.62	0.69	0.71	0.65	0.70
36	BLESS18R	Occupation	0.70	0.66	0.68	0.71	0.67	0.70
37	SMMSE6R	Apple, table,	0.62	0.53	0.66	0.62	0.52	0.65
38	SMMSE15R	Apple, table,	0.71	0.71	0.64	0.71	0.72	0.65
39	MSQ9R	Current president	0.81	0.82	0.73	0.81	0.82	0.72
40	MSQ10R	Previous presi-	0.74	0.74	0.67	0.73	0.74	0.66
41	BLESS21R	Dates for WW1	0.70	0.64	0.71	0.70	0.63	0.71
42	BLESS22R	Dates for WW2	0.78	0.74	0.77	0.78	0.73	0.76
43	BLESS23R	List months back-	0.78	0.71	0.78	0.77	0.7	0.78
44	BLESS24R	Count from 1 to 20	0.81	0.68	0.85	0.84	0.71	0.87
45	SMMSE10AR	Serial 7s	0.86	0.61	0.9	0.84	0.57	0.88
46	BLESS25R	Count backwards	0.91	0.66	0.94	0.91	0.62	0.94
47	S82R	Point to the quar-	0.73	0.61	0.77	0.78	0.66	0.81
48	S83R	How would you	0.79	0.69	0.81	0.80	0.69	0.83
49	S70R	Verbal repetition	0.68	0.59	0.72	0.67	0.58	0.71
50	BLESS26R	Recall John	0.69	0.69	0.63	0.69	0.69	0.63
	Model fit statistics:		0.930	0.962		0.929	0.964	

NOTE on the items of the measures: BLESS – The Blessed Memory-Information-Concentration Test; CMSQ - The CARE (Comprehensive Assessment and Referral Evaluation) Mental Status Questionnaire; MSQ – Mental Status Questionnaire; SMMSE – The Standardized Mini-Mental State Examination; SPMSQ – The Short Portable Mental Status Questionnaire; S - The INCARE (institutional version of the CARE) questionnaire.

Table 3
Cognition Item Pool: Internal Consistency Estimates

Sample	*N*	Ordinal Alpha	McDonald's	Explained Common
Total sample	3,504	0.984	0.984	83.608
Age				
≤ 65	224	0.979	0.980	77.699
66 – 75	449	0.985	0.985	85.479
76 – 85	1,235	0.982	0.982	83.702
86 – 95	1,412	0.984	0.983	84.493
> 95	163	0.982	0.982	77.589
Education				
No education	474	0.968	0.969	66.647
Grades 1-8	527	0.980	0.980	82.108
Grades 9 – 12	1,038	0.985	0.985	86.697
More than high school	558	0.986	0.986	85.862
Race				
Black	469	0.982	0.982	86.524
Hispanic	186	0.975	0.975	66.498
White	2,773	0.985	0.985	83.364
Sex				
Male	923	0.981	0.981	81.140
Female	2,581	0.984	0.985	83.819

Local independence

Because the $\chi 2$ statistics inflate in large samples, a 5 % random sample (n = 346) was extracted to calculate the LD statistics. In general, the local dependency statistics were in the range not indicative of collinearity (not shown). Nevertheless, there were a few item pairs with higher LD statistics that were over the threshold of 10.0. Details are available from the authors.

Reliability estimates

The classical test theory Cronbach's alpha estimated by IRTPRO was 0.95 (based on the listwise deletion of missing data) for the total sample. The corrected item-total correlations ranged from 0.42 to 0.71. The reliability estimates calculated in R were high. The McDonald's omega total was 0.984, the ordinal alpha was 0.984, both based on the polychoric correlations (Table 3). Estimates of reliability across age, education, race, and sex subgroups were high with none below 0.96. The reliability estimates (precision) at points along the latent trait (theta) estimated by IRTPRO were high for the total sample, ranging from 0.76 to 0.98 with an average of 0.94, with all but 3 estimates in the nineties.

IRT Parameter Estimates, Tests of DIF and Assessment of Magnitude and Impact

The graded response item parameters and their standard errors for the total sample are presented in Table 4. Overall, the a parameters were within the normal range. Only five items evidenced a values below 1.5 indicating lower discrimination. They were: Item (25) SMMSE7 – "No ifs, ands, or buts" (1.20); item (30) SMMSE12 – "Folding paper" (1.37); item (32) SMMSE14 – "Drawing pentagon" (1.25); item (34) BLESS16a – "How much schooling" (1.31); item (37) SMMSE6 – "Apple, table, penny (repeat)" (1.34). The most discriminating was item (18) MSQ5 – "Current year" (a = 3.07).

Forty-one items had 3 response categories while nine had up to 7. The difficulty 'b' parameters for the first response category were > 0 for 18 items where lower 'b' values indicate a more difficult item. The item with the lowest 'b' for the first response category was item (50) BLESS26 - "Recall John Brown's address" (-2.16) followed by the item (38) SMMSE15 – Apple, table, penny (recall)" (-1.64) and the item (45) SMMSE10A – "Serial 7s" (-1.52). The least difficult items were (33) BLESS16 – "Father's first name" (2.25) and item (26) SMMSE8 – "Name pencil" (2.12).

Table 4
Cognition Item Pool: IRT Parameter Estimates for the Total Sample (IRTPRO; n = 6,921)

Item	Item label	Mean	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*b6*
1	Spelling name	0.11	1.91	1.53	2.99				
2*	Age	0.49	2.02	-0.02	2.85				
3*	Month of birth	0.12	1.97	1.58	2.96				
4*	Year of birth	0.25	1.79	0.90	2.96				
5*	Birthplace	0.09	1.62	1.99	3.40				
6*	Where are we now	0.40	2.38	0.35	2.03				
7*	What do people do here	0.32	2.26	0.59	2.26				
8*	What is the name of this place	0.47	2.50	0.05	2.35				
9*	What floor are we on	0.28	2.29	0.67	2.53				
10	Street address of this place	0.69	2.11	-0.67	2.51				
11*	What city (town) is it	0.34	1.63	0.58	2.93				
12*	What county are we in	0.47	1.73	0.05	2.69				
13*	What state are we in	0.18	1.99	1.14	2.69				
14*	How long at this address	0.59	1.57	-0.35	2.81				
15*	Current month	0.43	2.72	0.17	2.25				
16*	Today's date	0.69	2.25	-0.62	2.42				
17*	Day of the week	0.47	2.41	0.08	2.36				
18*	Current year	0.51	3.07	-0.05	2.06				
19*	Current season	0.37	2.25	0.36	2.17				
20*	Hour of the day	0.38	2.18	0.35	2.21				
21*	Part of the day	0.29	2.29	0.65	2.17				
22*	Test of touching ears	0.65	1.78	0.38	0.94	1.67	2.61		
23*	Recall interviewer's name	0.59	1.62	-0.33	3.75				
24*	Interviewer's name (2nd try)	0.43	1.99	0.16	3.27				
25	No ifs, ands, or buts	0.45	1.20	0.12	3.38				
26*	Name pencil	0.06	1.99	2.12	3.29				
27*	Name wristwatch	0.09	2.24	1.74	2.92				

Item	Item label	Mean	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*b6*
28	Spell world backwards	2.14	1.89	-0.51	-0.23	0.06	0.39	0.69	2.25
29*	Close your eyes	0.24	1.87	0.88	2.56				
30*	Folding paper	0.77	1.37	-0.05	1.10	2.09	3.43		
31	Write sentence	0.30	1.91	0.57	2.39				
32	Drawing pentagon	0.74	1.25	-1.17	3.31				
33*	Father's first name	0.07	1.50	2.25	3.83				
34*	How much schooling	0.22	1.31	1.20	3.38				
35	Name of a school at-	0.23	1.62	1.01	2.71				
36*	Occupation	0.20	1.69	1.23	2.77				
37*	Apple, table, penny (re-	0.34	1.34	1.19	1.83	2.49	3.83		
38*	Apple, table, penny (re-	2.24	1.90	-1.64	-0.92	-0.42	2.69		
39*	Current president	0.48	2.52	-0.01	2.39				
40*	Previous president	0.71	2.02	-0.70	2.51				
41	Dates for WW1	0.58	1.64	-0.40	2.59				
42	Dates for WW2	0.59	2.18	-0.36	2.21				
43*	List months backwards	1.34	2.20	-0.75	-0.45	1.79			
44*	Count from 1 to 20	0.16	2.29	1.18	2.43				
45	Serial 7s	3.66	2.08	-1.52	-1.22	-1.02	-0.78	-0.4	2.25
46*	Count backwards from	0.46	2.37	0.02	2.24				
47*	Point to the quarter	0.09	1.76	1.80	3.24				
48*	How would you make 31	0.26	2.12	0.70	2.63				
49	Verbal repetition of the	1.10	1.51	-0.26	0.36	0.78	2.79		
50	Recall John Brown's ad-	4.19	2.02	-2.16	-1.83	-1.45	-1.07	-0.70	2.43

a = item discrimination; *b* = item difficulty; *= item selected for the final item bank
NOTE on the items of the measures: BLESS – The Blessed Memory-Information-Concentration Test; CMSQ - The CARE (Comprehensive Assessment and Referral Evaluation) Mental Status Questionnaire; MSQ – Mental Status Questionnaire; SMMSE – The Standardized Mini-Mental State Examination; SPMSQ – The Short Portable Mental Status Questionnaire; S - The INCARE (institutional version of the CARE) questionnaire.

DIF Results

Summary DIF results are presented in Table 5; statistical details for the demographic group comparisons are available from the authors. Thirty-eight items showed DIF with a minimum of one significant test; however, the number of items with two significant tests was reduced to 19 items. There were only eight items with all three criteria significant: item (1) BLESS1 = "Spelling name" where the item was more difficult for Hispanics; item (10) MSQ2 – "Street address for this place" with the item more difficult for patients with no education; item (25) SMMSE7 – "No ifs, ands, or buts" where the DIF was non-uniform, and less discriminant and easier overall for Hispanics; item (28) SMMSE10 – "Spell world backwards" where the item was less difficult for patients age 86 to 95 and more difficult for Hispanics; item (35) BLESS17 – "Name of school attended" evidenced non-uniform DIF and was more difficult overall for Hispanics; item (41) BLESS21 – "Date for WW1" where the item evidenced non-uniform DIF for the age groups 86 to 95 and age greater than 95, in the direction that the item was more discriminating but less difficult for both groups; Items (41) BLESS21 – "Date for WW1" and (42) BLESS22 – "Date for WW2" were both more difficult for the Hispanic patients and females; and, item (49) S70 – "Verbal repetition of the John Brown's address" which was more difficult for Hispanics.

Aggregate Impact. There was no scale level impact for age, education, race/ethnicity or sex DIF. All group curves were overlapping for all comparisons.

Information

Information functions were examined. Both item and scale information functions were bimodal with the highest information concentrating at upper levels of theta, i.e., at higher cognitive impairment, as well as at theta level 0.0 (see Table 6 for information functions and IRT reliability estimates). The highest information for the five most informative items were: item (18) MSQ5 – "Current year" (2.36 at theta level 0.0 and 2.35 at theta 2.0); item (15) MSQ4 – "Current month" (1.78 at theta 2.4 and 1.76 at theta 0.0); (39) MSQ9 – "Current president" (1.59 at theta levels 0.0 and 2.4); (8) S23 – "What is the name of this place?" (1.57 at theta levels 0.0 and 2.4); and item (17) BLESS12 – "Day of the week" (1.46 at theta 2.4 and 1.45 at theta 0.0). The five least informative items were: item (25) SMMSE7 – "No ifs, ands, or buts" (0.36 at theta levels 0.0 and 0.4); item (32) SMMSE14 – "Drawing pentagons" (0.39 at theta level -1.2); item (34) BLESS16A – "How much schooling?" (0.45 at theta levels 1.2 to 1.6 and 2.8); (37) SMMSE6 – "Apple, table, penny (repeat)" (0.57 at theta levels 2.0 to 2.4); (30) SMMSE12 – "Folding paper" (0.58 at theta levels 1.2 to 2.0). The item (25) SMMSE7 – "No ifs, ands, or buts" was the only one that exhibited DIF by all three estimators and item (30) SMMSE12 – "Folding paper" showed no DIF.

Table 5
Cognition Item Pool: IRT DIF Summary - Items with Significant DIF After Adjustment for Multiple Comparisons by Demographic Groups

Item #	Item name	Age (Reference group: Age 66 to 76; *n* = 6,879)				Education (Reference group: Grades 9 to 12; *n* = 4,885)			Race / Ethnicity (Reference group: Whites; *n* = 6,723)		Sex (Reference group: Female; *n* = 6,914)
		≤ 65	76 to 85	86 to 95	> 95	No educat.	Grades 1 to 8	> HS	Black	Hispanic	
1	BLESS1R -									U More;	
2	MSQ6R –										
3	MSQ7R -										
4	MSQ8R -										U
5	SPMSQ2R				U						
6	MSQ1R -		U	U		NU			U		
7	S22R -			U		NU		U	U		
8	S23R -										
9	S23AR -										
1	MSQ2R -		U	U	U	U			U	U Less	
1	BLESS8R -			U					U		
1	SMMSE4R			U		U					
1	S24CR -									U More;	
1	BLESS7R -					NU			U		
1	MSQ4R -		U	U					U		
1	MSQ3R -		U	U					U	U Less	
1	BLESS12R			U			U		U		
1	MSQ5R -		U	U					U	U Less	
1	BLESS15R									U More;	
2	BLESS14R			U							
2	BLESS13R										
2	S57R - Test		T1	U	T					U More;	
2	S60R - Re-										
2	CMSQ10R -										
2	SMMSE7R									NU Less	
2	SMMSE8R										NU
2	SMMSE9R										
2	SMMSE10		T1	U	U		U	T1	U	U More;	U
2	SMMSE11					U					

Item #	Item name	Age (Reference group: Age 66 to 76; *n* = 6,879)				Education (Reference group: Grades 9 to 12; *n* = 4,885)			Race / Ethnicity (Reference group: Whites; *n* = 6,723)		Sex (Reference group: Female; *n* = 6,914)
		≤ 65	76 to 85	86 to 95	> 95	No educat.	Grades 1 to 8	> HS	Black	Hispanic	
3	SMMSE12										
3	SMMSE13		U	U	U		U		U	U More	U
3	SMMSE14					U			NU		
3	BLESS16R										
3	BLSS16AR					U					
3	BLESS17R			U			NU			NU Less	U
3	BLESS18R						U				
3	SMMSE6R			U	U				U	T1	
3	SMMSE15	T1	T1	U	T		T1		U	T1	
3	MSQ9R -			U	T				U	U More	U
4	MSQ10R -			NU	N	U			U		U
4	BLESS21R	U	U	NU	N			U		U More;	U
4	BLESS22R	U		U	N			U		U More;	U
4	BLESS23R										
4	BLESS24R										
4	SMMSE10	T1	T1	U	U		T1	U	U	U More;	U
4	BLESS25R								U		
4	S82R -										
4	S83R - How										
4	S70R - Ver-	NU					U	T1		U More;	
5	BLESS26R	T1	T1	U	T				U	T1	T1

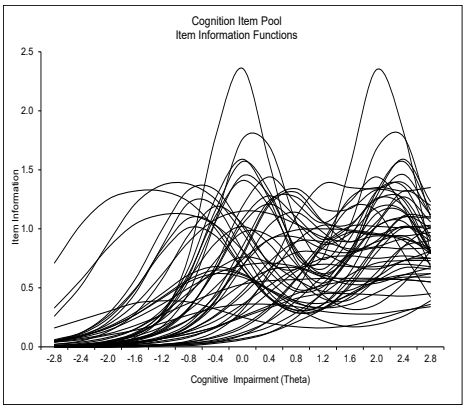
U = Uniform DIF; NU = Non-uniform DIF; discr. = discriminating; educat. = education; HS = high school; More/Less = More/Less likely to err

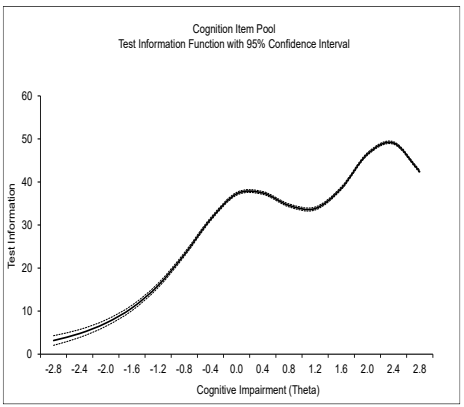
NOTE on the items of the measures: BLESS – The Blessed Memory-Information-Concentration Test; CMSQ - The CARE (Comprehensive Assessment and Referral Evaluation) Mental Status Questionnaire; MSQ – Mental Status Questionnaire; SMMSE – The Standardized Mini-Mental State Examination; SPMSQ – The Short Portable Mental Status Questionnaire; S - The INCARE (institutional version of the CARE) questionnaire.

Table 6
Cognition Item Pool: Item Response Theory (IRT) Reliability Estimates and Information Functions (IRTPRO; n = 6,921)

Theta	Reliability
-2.8	0.76
-2.4	0.83
-2.0	0.88
-1.6	0.91
-1.2	0.94
-0.8	0.96
-0.4	0.97
0.0	0.98
0.4	0.98
0.8	0.97
1.2	0.97
1.6	0.98
2.0	0.98
2.4	0.98
2.8	0.98
Overall (average)	0.94

Note: Reliability estimates are calculated for theta levels for which there are respondents





Correlations Among the Scales, Sum Scores and Theta Estimates

Correlations with the theta estimate (not shown) were high and ranged from 0.882 (CMSQ) to 0.913 (BLESSED). The theta estimate correlated 0.926 with the sum score; the intercorrelations among scales ranged from 0.830 to 0.942.

Discussion

A cognitive function item bank was developed for use in screening for cognitive impairment across stages, including in advanced dementia. The bank is intended for use among residents in long term services and support (LTSS) settings, including home care, day care, independent living, assisted living and nursing homes. Using latent variable models, including item response theory and factor analytic methods, we reduced a 50-item pool down to a 40-item unidimensional item bank with minimal differential item functioning by age, sex, race/ethnicity, and educational level. Two other items were removed because of intellectual property issues, yielding a final bank of 38 items. (Contact the authors for item sources.)

The item "season" has evidenced ethnicity- (Latino/ non-Latino subjects; Escobar et al., 1986; Marshall et al., 1997; Valle et al., 1991) and education- DIF (Escobar et al.; Jones & Gallo, 2002). Similarly, the "state recall" item has showed diverging difficulty associated with ethnicity (i.e., more difficult for Latinos [vs. non]), and poor discrimination for individuals at lower education levels (Teresi et al., 1995). The repeat/recall 3 objects item has been found to have education (Teresi et al.) - and ethnicity- DIF (easier for Latinos; Jones & Gallo) while the "spelling world backwards" item has demonstrated to be of higher difficulty for those participants with low education (Escobar et al.; Jones & Gallo) and for Latinos/as (Escobar et al.; Hohl et al., 1999). "No ifs, ands, or buts" has shown higher difficulty (Jones & Gallo) and poor discrimination for participants with relatively low levels of education (Teresi et al.), and less difficulty for Latinos/as (Escobar et al.; Teresi et al., Marshall et al.). The "does not close your eyes" item has evidenced education-DIF (Escobar et al.; Teresi et al.) and poor discrimination on race/ethnicity (Black and Latino participants; Teresi et al.). The "serial 7s" item has performed poorly and has exhibited ethnicity/racial- (higher difficulty for Latinos/as vs. non-Latinos/as (Escobar et al.; Hohl et al.) and for Latino/as and Blacks vs. White participants; Teresi et al.); and education-DIF (a poor discriminator- [Teresi et al.] and of higher difficulty for those with low educational level [Escobar et al; Jones & Gallo]). The "sentence completion" item has evidenced race- (lower difficulty for Whites as contrasted with Blacks), and ethnicity-DIF (less discrimination of Spanish-speakers; Marshall et al.), and education-DIF (higher difficulty for the those with lower education level; Escobar et al; Jones & Gallo; Teresi et al.). The "copy pentagons design" item has shown education-DIF whereas difficulty was determined to be higher for those with low educational level (Escobar et al; Jones & Gallo). A confounding effect of race/ethnicity and education in MMSE performance has been suggested (Gibbons et al., 2011; Ramirez et al., 2006).

Based on these analyses and previous studies, items with low information and DIF evidenced by all criteria are recommended for removal. These items include: (1) BLESS1 = "Spelling name"; (2) MSQ2 – "Street address for this place"; (3) SMMSE7 – "No ifs, ands, or buts"; (4) SMMSE10 – "Spell world backwards"; (5) BLESS17 – "Name of school attended"; (6) BLESS21 – "Date for WW1"; (7) BLESS22 – "Date for WW2"; (8) S70 – "Verbal repetition of John Brown's address" (8 items with

significant DIF). The item "No ifs, ands, or buts" is also the item with the lowest information and has evidenced DIF in numerous analyses. The suggestion to also exclude is the next least informative item (9) SMMSE14 – "Drawing pentagons". Because the item (10) "serial 7's" has been found in the literature to show DIF across several socio-demographic groups it is also recommended for removal. Finally, the item (11) requiring "recall of John Brown" is recommended for removal because of its link to the item calling for "repetition of John Brown". Additionally, an item associated with intellectual property concerns and which requires physical function capability is also recommended for exclusion: (12) "writing a sentence".

Although not identified here as problematic, several other items were flagged in the literature reviewed as showing DIF: season of the year, recall of the state, repeating three objects, sentence completion and following the command to close eyes. However, the evidence for DIF for these items was less conclusive. Thus, the final 38-item item bank is recommended for future work constructing computerized adaptive tests and short-forms for use in LTSS settings. As reviewed above, items with a high magnitude of DIF may impact the overall score and result in biased cognitive assessments. It is thus recommended that investigators avoid such items in developing short-forms and providing items for selection in CATs. The extensive information provided here provides a basis for developing a cognitive screening CAT framework for LTSS that may be less biased across diverse groups.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. https://doi.org/10.1080/10705510903008204

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Blessed, G., Tomlinson, B., & Roth, M. (1968). The association between quantitative measures of dementia and senile changes in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry, 114,* 797-811.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., & Rose, M., on behalf of the PROMIS Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS). Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J., Gershon, R., Hahn, E., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., & Hays, R. on behalf of the PROMIS Cooperative Group. (2010) The

Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005 – 2008. *Journal of Clinical Epidemiology,* 63(11), 1179-1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.

Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika, 83,* 696-732. https://doi.org/10.1007/s11336-018-9626-9

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: improved differential test statistics that account for sampling variability. *Educational and Psychological Measurement, 76,* 114– 140. https://doi.org/10.1177/0013164415584576

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. https://doi.org/10.2307/1165285

Cheville, A. L., Wang C., Yost, K. J., Teresi, J. A., Ramirez, M., Ocepek-Welikson, K., Ni, P., Marfeo, E., Keeney, T., Basford, J. R., & Weiss, D. J. (2021). Improving the delivery of function-directed care during acute hospitalizations: methods to develop and validate the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT). *Archives of Rehabilitation Research and Clinical Translation*, *3*(2), 100112. https://doi.org/10.1016/j.arrct.2021.100112

Cohen, A. S., Kim, S.- H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350. https://doi.org/10.1177/014662169301700402

Copeland, J., Kelleher, J., Kellett, J., Gourlay, A., Gurland, B., Fleiss, J., & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule, I, Development and reliability. *Psychological Medicine*, *6*, 439-449.

Crane, P., van Belle, G., & Larson, E. B. (2004). Teat bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Dubbelman, M. A., Jutten, R. J., Tomaszewski Farias, S. E., Amariglio, R. E., Buckley, R. F., Visser, P. J., Rentz, D. M., Johnson, K. A., Properzi, M. J., Schultz, A., Donovan, N., Gatchell, J. R., Teunissen, C. E., Van Berckel, B. N. M., Van der Flier, W. M., Sperling, R. A., Papp, K. V., Scheltens, P., Marshall, G. A., . . . Alzheimer Disease Neuroimaging Initiative, for the Alzheimer Disease Neuroimaging Initiative, National Alzheimer's Coordinating Center, the Harvard Aging Brain Study, the Alzheimer Dementia Cohort. (2020). Decline in cognitively complex everyday activities accelerates along the Alzheimer's disease continuum**.** *Alzheimer's Research & Therapy*, 12, 138.

Escobar, J. L., Burnam, A., Karno, M., Forsythe, A., Landsverk, J., & Golding, J. M. (1986). Use of the Mini-Mental State Examination (MMSE) in a community population of mixed ethnicity. *Journal of Nervous and Mental Disease, 174*, 607-614. https://doi.org/10.1097/00005053-198610000-00005.

Fieo, R., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Cella, D., & Teresi, J. A. (2016). Measurement Equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) Applied cognition - general concerns, short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58*(2), 255-307.

Filshtein, T., Chan, M., Mungas, D., Whitmer, R., Fletcher, E., DeCarli, C., & Farias S. (2020). Differential item functioning of the Everyday Cognition (ECog) Scales in relation to racial/ethnic groups. *Journal of the International Neuropsychological Society, 26*(5), 515-526. https://doi.org/10.1017/S1355617719001437

Fleer, P. F. (1993). *A Monte Carlo assessment of a new measure of item and test bias*. [Doctoral Dissertation, Illinois Institute of Technology]. Dissertation Abstracts International, 54(04B), 2266. ProQuest Dissertations Publishing. 9324223.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*(4), 309-326. https://doi.org/10.1177/01466219922031437

Folstein, M. F., Folstein, S. E. & McHugh, P. R. (1975). Mini-Mental State: A practical guide for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Fox, R. S., Zhang, M., Amagai, S., Bassard, A., Dworak, E. M., Han, Y. C., Kassanits, J., Miller, C. H., Nowinski, C. J., Giella, A. K., Stoeger, J. N., Swantek, K., Hook, J. N., & Gershon, R. C. (2022). Uses of the NIH Toolbox® in clinical samples. *Neurology, Clinical Practice, 12*(4), 307-319. https://doi.org/10.1212/CPJ.0000000000200060

Gershon, R. C., Lai, J. S., Bode, R., Choi, S., Moy, C., Bleck, T., Miller, D., Peterman, A., & Cella, D. (2012). Neuro-QOL: quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing. *Quality of Life Research, 21*(3), 475-486. https://doi.org/10.1007/s11136-011-9958-8

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH Toolbox for assessment of neurological and behavioral function. *Neurology, 80*(11 Supplement 3), S2-S6. https://doi.org/10.1212/WNL.0b013e3182872e5f

Gibbons, L. E., Crane, P. K., Mehta, K. M., Pedraza, O., Tang, Y., Manly, J. J., Narasimhalu, K., Teresi, J., Jones, R. N., & Mungas, D. (2011). Multiple, correlated covariates associated with differential item functioning (DIF): Accounting for language DIF when education levels differ across languages. *Ageing Research*, *2*(1), 19-25. https://doi.org/10.4081/ar.2011.e4

Golden, R. Teresi, J., & Gurland, B. (1983). Detection of dementia and depression cases with the Comprehensive Assessment and Referral Evaluation interview schedule. *International Journal of Aging and Human Development*, *16*(4), 242-254.

Golden, R., Teresi, J., & Gurland, B. (1984). Development of indicator-scales for the Comprehensive Assessment and Referral Evaluation interview schedule. *Journal of Gerontology*, *39* (2), 138-146.

Gurland, B., Fleiss, J., Goldberg, K., Sharpe, L., Copeland, J., Kelleher, M., & Kellet, J. A. (1976). Semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule, II, A Factor Analysis. *Psychiatric Medicine*, *6*, 451-459.

Gurland, B., Golden, R., Teresi, J., & Challop, J. (1984). The Short Care: An efficient instrument for the assessment of depression, dementia and disability. *Journal of Gerontology*, *39*(2):166-169.

Gurland, B., Kuriansky, J., Sharpe, L., Simon, R., Stiller, P., & Birkett, P. (1977-78). The Comprehensive Assessment and Referral Evaluation (CARE): -- Rationale, development and reliability. *International Journal of Aging and Human Development*, *8*(1): 9-42.

Gurland, B. J., & Wilder, D. E. (1984). The CARE interview revisited: Development of an efficient, systematic clinical assessment. *Journal of Gerontology*, *39*, 129-137.

Gurland, B. J., Wilder, D., Cross, P., Teresi, J., & Barrett, V. W. (1992). Screening scales for dementia: Toward reconciliation of cross-cultural findings. *International Journal of Geriatric Psychiatry, 7*, 105-113.

Haley, S. M., Coster, W. J., Andres P. L., Ludlow, L. H., Ni, P., Bond, T. L. Y., Sinclair, S. J., & Jette, A. M. (2004), Activity outcome measurement for postacute care. *Medical Care, 42*(1 Suppl), I49-I61. https://doi.org/10.1097/01.mlr.0000103520.43902.6c.

Hohl, U., Grundman, M., Salmon, D. P., Thomas, R. G., & Thal, L. J. (1999). Mini-Mental State Examination and Mattis Dementia Rating Scale performance differs in Hispanic and non-Hispanic Alzheimer's disease patients. *Journal of the International Neuropsychological Society, 5*(4), 301–307. https://doi.org/10.1017/s1355617799544019

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *Journals of Gerontology, 57B*, P548–P558.

Kahn, R., Goldfarb, A., Pollack, M., & Gerber, I. E. (1960). The relationship of mental and physical status in institutionalized aged persons. *American Journal of Psychiatry*, *117,* 326-328. https://doi.org/10.1176/ajp.117.2.120

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58*, 79-98.

Langer, M. M. (2008). *A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* [Doctoral dissertation, University of North Carolina at Chapel Hill]. University of North Carolina at Chapel Hill library. http://search.lib.unc.edu/search?R=UNCb5878458.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum.

Mann, A. H., Wood, K., Cross, P., Gurland, B., Schreiber, P., & Haefner, H. (1984). Institutional care of the elderly: a comparison of the cities of New York and London and Mannheim. *Social Psychiatry, 19*, 1-6.

Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. *Psychology and Aging, 12*, 718 –725. https://doi.org/10.1037/0882-7974.12.4.718.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Lawrence Erlbaum Associates.

Meade, A., Lautenschlager, G., & Johnson, E. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455. https://doi.org/10.1177/0146621606297316

Molloy, D. W., Alemayehu, E., & Roberts, R. (1991). Reliability of a standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination. *American Journal of Psychiatry, 148*(1), 102-105.

Morales, L., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. (2006). Item and scale differential functioning of the Mini-Mental Status Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care, 44*(11, Suppl. 3), S143-S151. https://doi.org/10.1097/01.mlr.0000245141.70946.29.

Mungas, D., Widaman, K. F., Reed, B. R., & Tomaszewski Farias, S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology, 25*(2), 260-269. https://doi.org/10.1037/a0021090

Muthén, L. K., & Muthén, B. O. (1998-2017). *M-PLUS Users Guide. Eighth Edition.* Muthén and Muthén.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatric Society, 53,* 695-699.

Orlando-Edelen, M., Stucky, B., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Quality of Life Research, 24*(1), 95-103. https://doi.org/10.1007/s11136-013-0540-4.

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. *Medical Care, 44*, S134-S142.

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society*, *22*, 433-444.

Raju, N. S. (1999). *DFITP5: A Fortran program for calculating dichotomous DIF/DTF* [Computer program]. Illinois Institute of Technology.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368. https://doi.org/10.1177/014662169501900405

Ramírez, M., Teresi, J., Holmes, D., Gurland, B., & Lantigua, R. (2006). Differential item functioning (DIF) and the Mini-Mental Status Examination (MMSE): Overview, sample, and issues of translation. *Medical Care, 44*(11, Suppl. 3), S95-S106. https://doi.org/10.1097/01.mlr.0000245181.96133.db

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22-S31.

Reise, S. P., Moore, T. M., Haviland, M. G. (2010). Bi-factor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559. https://doi.org/10.1080/00223891.2010.496477

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. https://doi.org/10.1037/met0000045.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114. https://doi.org/10.1007/BF02290599

Seybert, J., & Stark, S. (2012). Iterative linking with the differential functioning of items and tests (DFIT) method: comparison of testwide and item parameter replication (IPR) critical values. *Applied Psychological Measurement, 36*(6), 494-515. https://doi.org/10.1177/0146621612445182

Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. https://doi.org/10.1007/s11336-008-9101-0

Stump, T., Monahan, P., & McHorney, C. (2005). Differential item functioning in the Short Portable Mental Status Questionnaire. *Research in Aging, 27*, 355–384.

Teresi, J. A. (2007). Commentary: Scaling the Mini-Mental State Examination using item response theory. *Journal of Clinical Epidemiology*, *60*(3), 256-259.

Teresi, J., Golden, R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly White, Afro-American, Hispanic, and high and low education subgroups. *Journal of Clinical Epidemiology, 48*, 473-483. https://doi.org/10.1016/0895-4356(94)00159-N

Teresi, J., Golden, R., & Gurland, B. (1984). Concurrent and predictive validity of indicator-scales developed for the Comprehensive Assessment and Referral Evaluation interview schedule. *Journal of Gerontology*, *39*(2), 158-165.

Teresi, J., Golden, R., Gurland, B., Wilder, D., & Bennett, R. (1984). Construct validity of indicator-scales developed for the Comprehensive Assessment and Referral Evaluation interview schedule. *Journal of Gerontology, 39*(2), 147-157.

Teresi, J. A., Holmes, D., Ramírez, M., Gurland, B. J., & Lantigua, R. (2001). Performance of cognitive tests among different racial/ethnic and education groups: Findings of differential item functioning and possible item bias. *Journal of Mental Health and Aging, 7*(1), 79-89.

Teresi, J., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683. https://doi.org/10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H

Teresi, J. A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., & Holmes, D. (2000). Applications of item response theory to the examination of the psychometric properties and differential item functioning of the CARE Dementia Diagnostic Scale among samples of Latino, African-American and White non-Latino elderly. *Research in Aging, 22*, 738-773. https://doi.org/10.1177/0164027500226007

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Kleinman, M., Wang, C., Weiss, D., & Cheville A. (2022). Challenges in measuring applied cognition: Measurement properties and equivalence of the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT) applied cognition item bank. *Archives of Physical Medicine and Rehabilitation, 103*(5S):S118-S139. https://doi.org/10.1016/J.APMR.2020.12.029

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging & Health, 24*(6), 1044-1076. doi:10.1177/0898264312436877.

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions. *Psychometrika*, *86*(3), 674–711. https://doi.org/10.1007/s11336-021-09775-0.

Thissen, D. (1991). *MULTILOGTM User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory.* Scientific Software, Inc.

Valle, R., Hough, R., Kolody, B., Cook-Gait, H., Velazquez, G. F., & Jimenez, R. (1991). *The validation of the Blessed Mental Status Test and the Mini-Mental State Examination with a Hispanic population. Final Report to the National Institute of Mental Health, The Hispanic Alzheimer's Research Project (HARP).* San Diego State University.

Wainer, H. (1993). Model-based standardization measurement of an item's differential impact. In P. W. Holland, & H. Wainer (Eds.). *Differential Item Functioning* (pp. 123-135). Lawrence Erlbaum, Inc.

Wang, W. -C., Shih, C. -L, & Sun, G. -W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*, 687-708. https://doi.org/10.1177/0013164411426157

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42-57. https://doi.org/10.1177/0146621607314044.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547. https://doi.org/10.1177/0013164412464875.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.