The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 2025, **80**(S1), gbae043
https://doi.org/10.1093/geronb/gbae043
Advance access publication 10 April 2024
Supplement Article: Studying Cognition in the National Social Life, Health, and Aging Project: Methodological and Analytic Considerations, Strategies and Findings From Round 4

Race/Ethnicity and the Measurement of Cognition in the National Social Life, Health, and Aging Project: Recommendations for Robustness

James Iveniuk, PhD,[1,*] Selena Zhong, MA,[1] Jocelyn Wilder, MPH,[1] Gillian L. Marshall, PhD,[1] Patricia Boyle, PhD,[2] Jennifer Hanis-Martin, PhD,[1] Louise Hawkley, PhD,[1] Lissette M. Piedra, PhD,[3] Alicia R. Riley, PhD,[4] and Haena Lee, PhD[5]

[1]The Bridge, and the Center on Equity Resarch, NORC at the University of Chicago, Chicago, Illinois, USA.
[2]RUSH Alzheimer's Disease Center, RUSH University, Chicago, Illinois, USA.
[3]Department of Latina/Latino Studies, School of Social Work, University of Illinois Urbana-Champaign, Urbana, Illinois, USA.
[4]Department of Sociology, Global and Community Health, University of California Santa Cruz, Santa Cruz, California, USA.
[5]Department of Sociology, Sungkyunkwan University, Seoul, Korea.

*Address correspondence to: James Iveniuk, PhD. E-mail: iveniuk-james@norc.org

Decision Editor: Hui Liu, PhD (Social Sciences Section)

Abstract

Objectives: In this study, we examine the measurement of cognition in different racial/ethnic groups to move toward a less biased and more inclusive set of measures for capturing cognitive change and decline in older adulthood.

Methods: We use data from Round 2 ($N = 3,377$) and Round 3 ($N = 4,777$) of the National Social Life, Health, and Aging Project (NSHAP) and examine the study's Survey Adjusted version of the Montreal Cognitive Assessment (MoCA-SA). We employ exploratory factor analyses to explore configural invariance by racial/ethnic group. Using modification indexes, 2-parameter item response theory models, and split-sample testing, we identify items that seem robust to bias by race. We test the predictive validity of the full (18-item) and short (4-item) MoCA-SAs using self-reported dementia diagnosis, instrumental activities of daily living, proxy reports of dementia, proxy reports of dementia-related death, and National Death Index reports of dementia-related death.

Results: We found that 4 measures out of the 18 used in NSHAP's MoCA-SA formed a scale that was more robust to racial bias. The shortened form predicted consequential outcomes as well as NSHAP's full MoCA-SA. The short form was also moderately correlated with the full form.

Discussion: Although sophisticated structural equation modeling techniques would be preferable for assuaging measurement invariance by race in NSHAP, the shortened form of the MoCA-SA provides a quick way for researchers to carry out robustness checks and to see if the disparities and associations by race they document are "real" or the product of artifactual bias.

Keywords: Alzheimers disease, Functional health status, Measurement, Mortality

Approximately one in 10 adults older than 65 years in the United States have dementia (Fuller-Thomson & Ahlin, 2022; GBD 2016 Dementia Collaborators, 2019; Manly et al., 2022), and an additional 22% have some form of mild cognitive impairment (MCI; Manly et al., 2022). The prevalence of cognitive impairment is worrisome because cognitive decline is associated with poor health and premature mortality in older adulthood (Luck et al., 2015; Tilvis et al., 2004). Given the importance of cognitive functioning, researchers are interested in understanding the determinants of good cognitive function and tracing the implications of poor cognitive function for a range of physical, psychological, and social outcomes (Baumgart et al., 2015; McSorley et al., 2019; Perry et al., 2022). Disparities in cognitive function by sociodemographic groups can also shed light on inequities in U.S. society. However, to properly examine disparities, it is essential to have accurate and unbiased measurements of cognition.

Cognition assessments are fraught because many psychometric measures have historically incorporated researchers' own racial prejudices and implicit biases, as well as structural and institutional biases (Kotwal et al., 2015; O'Driscoll & Shaikh, 2017; Reynolds et al., 2021; Shuttleworth-Edwards, 2016; van de Vijver & Tanzer, 2004). In establishing unbiased measures, special consideration should be given to U.S. minority aging populations, who tend to have less access to and consequently lower levels of education and income, and experienced systemic racism during their young adulthood and throughout their lives. Because the life experiences of minoritized groups differ considerably from those of Whites, even those from lower socioeconomic positions, cognitive tests may underestimate cognitive abilities in minority ethnic groups (Parker & Philp, 2004).

Received: May 4 2023; Editorial Decision Date: December 4 2023.

Furthermore, literal translations of recall words from English to other languages may not carry the same social significance and, thus, may be more or less difficult to remember. Accordingly, we develop and evaluate a cognition measure that can minimize potential biases by race in the National Social Life, Health, and Aging Project (NSHAP).

NSHAP measures cognition using a survey-adapted form of the Montreal Cognitive Assessment (MoCA-SA; Shega et al., 2014). The MoCA-SA is composed of 18 items that together form a scale that is used as a continuous measure of cognition. Low scores are thought to indicate MCI, with the specific threshold for MCI varying across settings; the lowest scores indicate serious cognitive impairment, such as dementia. However, some items in the MoCA-SA may be more sensitive to racial bias than others. In this article, we investigate the MoCA-SA for measurement invariance across racial and ethnic groups. Our overall objective is to identify a short form of the MoCA-SA that only includes the items that resist racial bias, which can be used in future research to assess racialized disparities in cognition with better accuracy.

Our analyses reveal that some, but not all, of the items that comprise the MoCA-SA function differently across racial and ethnic groups. We created a shortened version of the MoCA-SA that only included cognitive measures that functioned similarly across racial groups to reduce biases in the cognitive measure. We then tested whether the shortened MoCA-SA predicted consequential outcomes in similar ways to the full MoCA-SA to assess its validity. We then close with recommendations for researchers intending to use the MoCA-SA in their own analyses.

Cognition, Race, and Previous Assessments of the MoCA

Previous research found that belonging to a racial/ethnic group is a risk factor for poorer cognitive function in later life (Rajan et al., 2021). According to the Centers for Disease Control (CDC, 2016), among adults 65 years and older, African Americans have the highest prevalence of Alzheimer's disease and related dementias (13.8%), followed by Hispanics (12.2%), non-Hispanic Whites (10.3%), American Indian and Alaskan Natives (9.1%), and Asian and Pacific Islanders (8.4%). More recent evidence suggests that improvements in educational levels reduce the risk of cognitive decline (Larson et al., 2013). Yet, despite increases in educational attainment, some groups still have high rates of cognitive decline compared to others. For example, African Americans are about two times more likely than White Americans to have Alzheimer's and other dementias, and Hispanics are about one and one-half times more likely than Whites to have Alzheimer's disease and other dementias (Alzheimer's Association, 2020). Assessing measurement invariance is especially important for the study of cognitive aging in diverse samples, such as NSHAP, and for the study of racial inequities in cognitive decline and related health disparities.

With reference to the MoCA specifically and its applicability across racial groups, the literature provides justification for an investigation of the measure in NSHAP. Although one previous analysis of the MoCA-SA at Round 2 of NSHAP found no differential item functioning by race (Kotwal et al., 2015), this analysis did not run tests for measurement invariance. The original MoCA was developed in French based on the clinical intuitions of one of its authors and later when it was formally validated, identical versions were administered in French and English except for items used in the repetition task (Nasreddine et al., 2005). When the MoCA was later adapted to accommodate a survey environment, cognitive interviewing techniques (Willis, 2005) and direct observation of participant reaction were used to evaluate the clarity and wording of specific items and to probe the perceived meaning of phases (Shega et al., 2014). These interviews revealed modifications needed for implementing the MoCA in new survey contexts (usually the home) by nonmedical survey interviewers. These modifications included integration into computer assisted personal interviewing technology, administration by nonmedically trained field interviewers, and reducing average administration time to 12 min. In addition, changes were made to select English words to better resonate with an American English-speaking population and to facilitate comprehension. For example, Shega et al. (2014) observed that many participants reported difficulty understanding the directions to the trails-b task, which originally asked people to draw a line from a number to a letter in ascending order. Substituting the word "ascending" with "increasing" led to improved comprehension of the instructions and higher completion rates for the item. Although these interviews enhanced the MoCA-SA, these cognitive interviews did not over-sample participants from various ethnoracial groupings, and no cognitive interviews were conducted with participants with Spanish-speaking preferences. Thus, the potential for ethnoracial and linguistic disparities remained. Kotwal et al. (2015) subsequently showed that the MoCA-SA could be administered reliably in a survey setting and yield similar performance across different demographic subgroups. Yet, caveats remained, especially regarding education, which are consistent with other studies into educational discrepancies (Adana Díaz et al., 2021). Those with less than a high school education had more difficulty with the subtract 7s and fluency items, which seem to require more math and language skills (Kotwal et al., 2015). Given cohort effects related to legal segregation and an increase in migration from countries with few educational opportunities, many African American and Hispanic older adults have lower levels of education than the general White population (Adana Díaz et al., 2021).

Additionally, cross-cultural analyses of the MoCA reveal a literature with little harmonization and often poor research designs (O'Driscoll & Shaikh, 2017). The MoCA was validated in English (Nasreddine et al., 2005), which makes some cognitive tasks, such as the phonemic fluency, alternating trail-making, and sentence repetition tasks, difficult to translate across language and culture without rigorous translation protocols that include independent translations and cognitive interviews (Eremenco et al., 2005). Although many versions of the MoCA have been adapted to identify MCI in different geographic regions, there is a wide range of suggested cutoffs for MCI across these different versions (O'Driscoll & Shaikh, 2017). In other words, there is no inclusive adaptation of the MoCA to date—only different cutoffs for different populations. Given the difficulty in creating an inclusive and valid MoCA test, MoCA scores should be interpreted with caution when it comes to assessing cognitive differences across cultural and racial groups. Accordingly, we need to interpret the MoCA-SA in NSHAP cautiously because although the complete MoCA (and its limitations) is well understood, NSHAP's modified version has not received the same level of scrutiny. The

MoCA-SA was also validated for NSHAP with the majority population in mind, which is likely to mask underlying discrepancies related to education and by extension, race. Specifically, the goal of this article is to employ exploratory factor analyses to explore configural invariance by racial/ethnic group and use modification indexes, two-parameter item response theory (IRT) models, and split-sample testing to identify items that seem robust to bias by race.

Method

Data

NSHAP is a nationally representative longitudinal survey of community-dwelling older Americans that began in 2005–2006 (Round 1) with follow-up interviews conducted every 5 years. The NSHAP sample is constructed using a multistage area probability sample design with an oversampling of African Americans and Latinos to achieve a balanced sample (O'Muircheartaigh et al., 2009). The overall weighted response rate for Round 1 was 75.5% with 3,005 completed interviews; there was only a 4.5% difference in response rates between minority and nonminority households. Approximately 6% of the interviews were conducted in Spanish (O'Muircheartaigh et al., 2009, 2014). In 2010–2011, NSHAP reinterviewed the original respondents and added a sample of the original respondents' spouses and cohabiting partners and attempted to interview R1 nonrespondents (Round 2). The unconditional response rate in R2 was 73.9%, which is very close to R1's response rate of 75.5%, indicating very slight attrition, although the response rate was lower among the older age cohorts (O'Muircheartaigh et al., 2014). In 2015–2016, NSHAP conducted a third round of interviews with surviving respondents (Cohort 1) and recruited a cohort of Baby Boomers and their spouses/partners (Cohort 2). The conditional response rate of Cohort 1 was 91%, whereas the response rate of Cohort 2 was 76% (O'Muircheartaigh et al., 2009). The MoCA-SA was included only in Rounds 2 and 3. Details about recruitment and sample characteristics are available elsewhere (Lindau et al., 2007; O'Muircheartaigh et al., 2014; Waite et al., 2021). Data for this paper come from Rounds 2 and 3.

Measures

Cognition

Cognition was assessed in NSHAP using the MoCA-SA. The MoCA-SA contains 18 items: the modified trail making test-B, clock drawing test, naming a rhinoceros, forward and backwards digit span, serial 7 subtractions, sentence repetition, phonemic fluency (list words with the letter "F"), abstraction (similarity between a watch and a ruler), 5-word delayed recall, and temporal orientation (month and date). It is designed to assess executive function, language, working memory, episodic memory, temporal orientation, and visuospatial ability. All items are shown in Table 1. For all the cognitive items except the serial subtract 7s, a score of 1 was given for correct responses to each cognitive item and a score of 0 for incorrect responses. For the serial subtract 7s item, respondents were asked to start with 100 and subtract 7 and to keep counting down; up to 6 responses were recorded. Because more correct responses reflect better cognitive ability for this item, a score of 0 was given for no correct answers, a score of 1 for 1 correct answer, a score of 2 for 2–3 correct answers, and a score of 3 for 4–5 correct answers, resulting in a range of 0–3. For the MoCA-SA, there are two versions of the measure that are for use with NSHAP: one that includes a correction for educational level (i.e., different thresholds for people based on education) and one without this correction (Shega et al., 2014). We specifically use the version *without* this correction because our short form (i.e., the one more robust to racial bias) does not include any such correction and we aim to compare like to like as much as possible.

Table 1. Domains and Questions in NSHAP's Version of the MoCA

Domain	Question	Content of question
Orientation	Month	Give today's month and date
	Date	
Naming	Rhino	Asked to identify a picture of a rhinoceros
Visuospatial	Contour	Draw a clock and set the time to 10 after 11
	Numbers	
	Hands	
Executive function	Trails	Draw a line going from a number to a letter in increasing order
Memory	Face	Repeat 5 words after a delay
	Velvet	
	Church	
	Daisy	
	Red	
Attention	Forward	Repeat 5 digits in forward order
	Backward	Repeat 3 digits in backwards order
	Serial7	Subtract 7 from 100 and keep counting down by 7
Language	Fluency	List as many words as you can starting with the letter F
	Cat	Repeat a sentence
Abstraction	Watch/Ruler	How are a watch and ruler alike?

Note: MoCA = Montreal Cognitive Assessment; NSHAP = National Social Life, Health, and Aging Project.

To assess how well our shortened MoCA-SA predicted consequential outcomes, we included self-reported dementia diagnosis, instrumental activities of daily living (IADLs), proxy report of dementia-related death at Round 3, and dementia-related cause of death as recorded in the National Death Index (NDI) as dependent variables. For each IADL, respondents were given a score of 1 if they report any difficulty with each of the eight activities. The scores were summed into a scale with a range of 0–8; higher scores indicated more difficulty with IADLs. Mortality status at Round 3 was determined through proxy interviews; proxies reported the medical conditions that the respondent had prior to death, which we interpret as contributing causes. We linked NSHAP data to the NDI. For respondents who were deceased at Round 3, we examined their cause of death in the NDI. If vascular dementia, unspecified dementias, Alzheimer's disease, or other neurodegenerative diseases contributed to death, these cases were coded as 1 and all others were coded as zero (survivors and nonsurvivors). The International Classification of Diseases codes used for the NDI measure were F01, F03, G30, and G31; all subcodes were used as well. Covariates used in the study include age, gender, race/ethnicity (*non-Hispanic*

White; non-Hispanic Black; Hispanic; Else), and education (*years of schooling*).

Analysis

The analysis took place in four stages. First, we tested for different forms of measurement invariance. We tested for configural invariance by running exploratory factor analyses separately for each racial and ethnic group and examining whether factor structures differed between groups (loadings, eigenvalues, and variances). We tested for metric and scalar invariance by relaxing the assumption of equal loadings and equal thresholds and performing log-likelihood tests for each possible comparison (configural vs metric, configural vs scalar, scalar vs metric).

In the second stage, we sought to identify which items functioned differently across racial/ethnic groups using modification indexes (Whittaker, 2012). Modification indexes allow the analyst to detect which items would improve model fit if their loadings and thresholds were allowed to vary across groups. The index is calculated by first holding the thresholds and loadings as constrained to be equal across groups and then allowing the parameter of interest to vary across groups. We then calculate a chi-squared test, comparing the model-based covariance matrix to the actual covariance matrix for both the constrained and unconstrained versions of the model. The modification index is the difference between the two chi-squared tests, showing how much model fit would improve if we allowed this parameter to vary across ethnoracial groups. Higher modification indexes are indicators of greater variability of the parameter across groups. We chose a cutoff of 3.84 for whether or not we would consider a parameter, and its corresponding item, as contributing to poor model fit (i.e., by behaving differently across groups). We chose 3.84 as the threshold for significance at $p < .05$ for a chi-squared test with one degree of freedom because comparing the two models allows one extra parameter to vary for a specific group. Note that indicators were treated as dichotomous for this analysis and the one "ordinal" measure of cognition (serial 7's) was dichotomized. We note items with modification indexes lower than 3.84. We dropped items sequentially based on the sum of their modification indexes across racial groups as a proxy for their contributions to measurement variance. Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), and log-likelihood tests comparing each model in the sequence to the model that preceded it were calculated to assess improvements in model fit. We thereby arrive at a "select" group of items that appear to behave the same across groups.

In the third stage, we take these select measures and test them *again* using IRT models. We do this to ensure the highest possible sensitivity to measurement invariance by group. Briefly, IRT models nest questions within respondents, and attribute a "difficulty" to each question, as well as an "ability" parameter—allowing for differences in test-taking ability from person to person. From this, for each question we obtain a "difficulty" parameter as well as a "discrimination" parameter, which is how well the question discriminates between respondents *at the same ability levels*. For each select item, we fit models that first allow the item's difficulty and discrimination parameters to vary by ethnoracial group, and then test that model *against* a model that does not allow either to vary. If there is improved model fit, according to a standard likelihood ratio test, then we can say that the item behaves differently for different ethnoracial groups. We also perform a likelihood ratio test where the difficulty parameter alone varies by group, but the discrimination parameter is constrained by group. After arriving at a model where the items that "should" vary by group do in fact vary by group, we carry out split-sample testing, where we fit the arrived-at model on one random half of the sample and test how well that model's parameters fit the *second half of the sample*. If it does fit well, we can more plausibly say that the model's fit is not an accidental one. For our final model, we plot item-characteristic curves, for each item that varies across ethnoracial groups. These can be read as follows: On the *y*-axis is the probability of answering the item successfully; on the *x*-axis is latent test-taking ability (for this exercise, cognitive function). The curves show how well the item discriminates between people on the basis of latent cognitive function; if the curve is set further to the left, it is "easy," if it is set further to the right, it is "difficult"; if the curve is steep, it discriminates well because once a person reaches a certain level of ability, they can reliably get it right (Nguyen et al., 2014).

In the fourth and final stage, we calculate a short-form cognition measure based on the items that "survive" the rigorous tests above. Our entire process for arriving at this short-form measure is outlined in Figure 1, recapitulating what the text above says. With the short form in hand, we compare it to the long form measure through the following tests:

(1) Percent of variance explained in each measure by race/ethnicity alone; uses data from Round 3 of NSHAP (analytic sample of 4,607; MoCA forms at Round 3).
(2) Pearson correlations between the short and long forms for all groups, pooled and separately. We carry these out separately for Round 2 and Round 3 (analytic samples of 3,363 and 4,607, respectively; MoCA forms at Round 2 and Round 3, respectively).
(3) Test–retest Pearson correlations between Round 3 and Round 2 for all groups, pooled and separately (2,371 surviving from R2 to R3).
(4) Predictive validity of each measure for
   a. Respondent-reported dementia (Round 3; analytic sample of 4,607; key predictor is MoCA at Round 3, covariates at Round 3);
   b. Respondent-reported functional health problems (IADLs; Round 3; analytic sample of 4,607; key predictor is MoCA at Round 3, covariates at Round 3);

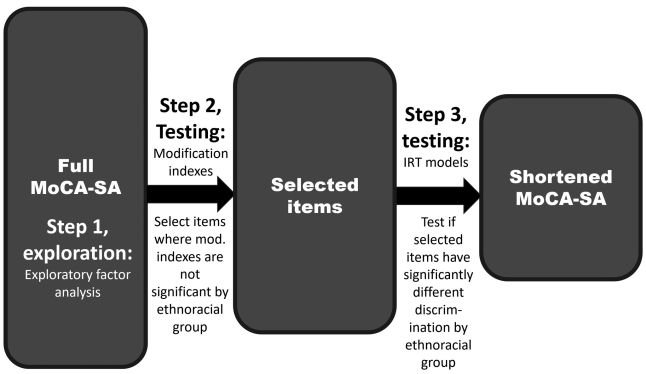

Figure 1. Process for selecting items for shorted Survey Adjusted version of the Montreal Cognitive Assessment (MoCA-SA), which will be robust to bias by ethnoracial group.

c. Dementia as a contributing cause of death, according to proxy reports (between Round 2 and Round 3; analytic sample of 3,363; key predictor is MoCA at Round 2, covariates at Round 2); and
d. Dementia as a key cause of death, according to data linked from the NDI (between Round 2 and Round 3; analytic sample of 3,363; key predictor is MoCA at Round 2, covariates at Round 2).

Note that the full R2 and R3 samples are 3,377 and 4,777, respectively; we restricted the four steps above to respondents who were aged 50 and older, thus the numbers above are less than the full sample. Sample sizes in regression tables will also be smaller, due to missing data. For the analyses under (4) above, item b can be assessed using ordinary least squares regression, controlling for education, gender, and age. Items a, c, and d represent dichotomous outcomes, and as we will see below, are relatively rare events in the NSHAP sample. Maximum likelihood estimation of logistic regression suffers from small-sample bias, and the degree of bias is strongly dependent upon the number of cases in the less-frequent of the two categories, or the "event" or "nonevent" (Allison, 2012). It also increases the possibility of separation, either complete (a perfect association between two dichotomous variables) or quasi-complete (one cell in the table is empty). In both cases, the estimates will be distorted and the model may not even converge (King & Zeng, 2001). Firth's penalized likelihood estimation function addresses both these issues and is far less computationally demanding than other methods such as "exact" logistic regression (Firth, 1993). It accomplishes this by introducing a "penalty" factor into standard maximum likelihood estimates that corrects for the expected bias arising from rare events. We therefore employ Firth penalized likelihood estimation models for these outcomes. Cognition scores were standardized before being entered into the model.

Comparison of logistic regression coefficients across models is challenging due to the well-known problem of assuming equal error variance across models for the key predictor. Because we assume the short form and long form of the MoCA-SA explain different degrees of variance in every outcome, we have cause to worry about this issue. We therefore compare predicted probabilities at the mean for each measure and in increments of 0.5 standard deviations (to a minimum of −1.5 and a maximum of 1.5). All other predictors were held at their mean when calculating predicted probabilities.

Exploratory factor analysis and all third-stage analyses were carried out in Stata v. 17, using the "factor," "regress," "firthlogit," "irt," and "margins" commands (StataCorp, 2021). Modification indexes and tests of measurement invariance were calculated in Mplus version 8 (Muthén & Muthén, 2017).

Results

Table 2 shows results for the exploratory factor analyses, first for the entire Round 3 sample and then by ethnoracial group. We show factor loadings, eigenvalues, and proportions of variance explained by the first unrotated factor. We can see for the entire sample, as well as for non-Hispanic Whites and Hispanics, that the first unrotated factor soaks up a large amount of the shared variance (greater than 95%). However, for non-Hispanic Black respondents, less of the variance is captured by a single factor (approximately 85%), and the second unrotated factor (not shown) captures 29% of the variance and has an eigenvalue of 0.85; it does not rise to the level of consideration as a second factor, using the conventional criterion of eigenvalues greater than or equal to 1, and so we still consider cognition to be a unidimensional latent construct for this group. For "else" there is much greater heterogeneity in factor structure (only two-thirds of the shared variance is

Table 2. Step 1; Exploratory Factor Analysis; Factor Loadings From First Unrotated Factor. Round 3 Data (N = 4,777)

Domain	Question	All	Non-Hispanic White	Non-Hispanic Black	Hispanic	Else
Orientation	Month	0.243	0.217	0.212	0.325	0.278
	Date	0.306	0.329	0.252	0.325	0.378
Naming	Rhino	0.400	0.344	0.432	0.375	0.371
Visuospatial	Contour	0.163	0.179	0.124	0.160	0.097
	Numbers	0.408	0.340	0.375	0.481	0.392
	Hands	0.407	0.371	0.326	0.439	0.440
Executive function	Trails	0.514	0.448	0.433	0.534	0.580
Memory	Face	0.384	0.386	0.365	0.457	0.313
	Velvet	0.501	0.474	0.488	0.454	0.571
	Church	0.418	0.436	0.350	0.501	0.368
	Daisy	0.409	0.410	0.385	0.457	0.288
	Red	0.454	0.452	0.466	0.493	0.329
Attention	Forward	0.296	0.203	0.210	0.353	0.339
	Backward	0.412	0.336	0.390	0.400	0.431
	Serial 7	0.571	0.483	0.563	0.578	0.526
Language	Fluency	0.441	0.384	0.489	0.423	0.448
	Cat	0.320	0.246	0.320	0.409	0.386
Abstraction	Watch/Ruler	0.281	0.220	0.252	0.189	0.166
Eigenvalue		2.841	2.339	2.522	3.204	2.756
Prop. variance explained		0.998	0.999	0.856	0.959	0.677

captured by the first factor) and the second factor (not shown) has an eigenvalue of 1.09 (28% of variance explained by the second factor). This suggests a more complex factor structure could exist for the "else" group. However, note that we still treat this as unidimensional given the small sample size for the "else" group ($n = 179$) and the heterogeneous nature of this group. We have not yet done any significance testing, however, at a glance, we can observe some heterogeneity in factor loadings (e.g., F-fluency and "rhino" are more strongly correlated with the latent factor, according to their loadings, for Hispanics compared to other groups).

Table 3 provides results from investigating specific MoCA-SA items using modification indexes—cells in the table are left blank if the index fell below the threshold of 3.84, and the rightmost column sums the modification index scores for each item. There were no items for White respondents that exceeded the 3.84 cutoff, and thus, their column is entirely absent. Some items such as counting forward show huge modification indexes in both their thresholds and loadings for both Black and Hispanic respondents. F-fluency is also particularly problematic for both Black and Hispanic groups. Trails also shows high modification indexes, especially for Black respondents. However, we can see that there is a subset of items that has no modification indexes at all above the 3.84 cutoff for any group. We dropped potentially problematic items, sequentially, in the order of their summed modification indexes (i.e., the rightmost column, from smallest to largest), and at each stage, we observed changes in the AIC, BIC, and log-likelihood tests comparing improvements in model fit with each item we dropped. The AIC and BIC steadily declined with each item we dropped and the log-likelihood test was always significant at $p < .001$. Thus, we retained *only* those items with a zero-sum score as our "select" measures for the IRT model testing.

Table 4 shows our IRT model tests of these select measures. We can see that in this new round of testing, only four items survive to become part of the short form: month of the year, and the three "clock draw" or "visuospatial" items. None of the items have comparable difficulties across the four groups (observe the very low p values in the last column), but this does not indicate that they are any worse at discriminating between people at the *same ability level* across the four groups. It only indicates that different groups have different rates of success with these items, but not that the items are any worse at detecting actual cognitive function across groups. Thus, we arrive at our four-item short form. Split-sample testing revealed very good model fit, according to BIC and AIC values. Testing a model fit to one random half of the Round 3 sample on the second random half, and comparing that model to one fit *directly* on the second random half revealed a reduction of AIC and BIC of less than 1%. We repeated this analysis 10 times on one random bifurcation of the sample and received similar results—less than a 1% decrease (i.e., improvement in model fit) each time.

In Figure 2, we show item response curves from the final model. We can see that for the "memory" items (face, velvet, church, daisy, and red), the stand-out group is the "else" category, for all but "velvet." The curves are steepest (i.e., discrimination based on test-taking ability is best) for "velvet," although the Hispanic group has a slightly flatter slope compared to the other groups (i.e., the test is less good at discriminating based on ability in this group). The curve is also shifted noticeably forward for White respondents on all memory items but "velvet," indicating that one needs to have a higher latent ability to get this item right, for Whites. The single "abstraction" item ("what do a watch and ruler have in common?") performs very poorly at detecting differences in ability for every group except Black respondents, where the curve is noticeably steeper. The "F-fluency" question also

Table 3. Step 2; Modification Indexes, Describing Differences in Item Functioning Across Racial/Ethnic Groups

Domain	Question	Black		Hispanic		Else		Sum
		Threshold	Loading	Threshold	Loading	Threshold	Loading	
Orientation	Month							0
	Date							0
Naming	Rhino			4.543		20.076		24.619
Visuospatial	Contour							0
	Numbers							0
	Hands							0
Executive function	Trails	12.345	12.782	7.443	9.439	5.498		47.507
Memory	Face					7.225		7.225
	Velvet					4.594		4.594
	Church	4.072	4.148		4.111			12.331
	Daisy	10.36	10.623	7.767	9.447			38.197
	Red							0
Attention	Forward	35.218	34.262	22.006	31.16			122.646
	Backward							0
	Serial 7	9.004	9.304	6.835	8.404			33.547
Language	Fluency	10.873	11.158	12.524	15.346	4.892		54.793
	Cat	4.037	4.111			27.188		35.336
Abstraction	Watch/Ruler					6.545		6.545

Table 4. Step 3 Part 1; IRT Tests of Select Items, Chosen From Items in Table 2 With Modification Index Values of Zero

Domain	Item	Likelihood ratio test *p* value	Likelihood ratio test *p* value, discrimination constrained
Orientation	Month	.522	.000
	Date	.000	.000
Visuospatial	Contour	.229	.000
	Numbers	.129	.000
	Hands	.446	.000
Memory	Red	.000	.000
Attention	Backward	.000	.000

Note: IRT = item response theory.

performs well in terms of detecting latent ability for Black respondents, but falls short for other groups, especially Hispanics. Regarding "cat" (correctly repeating the phrase "The cat always hid under the couch when dogs were in the room"), for non-Hispanic Whites, the item has very limited use at detecting differences in ability, but it does much better for all other groups. Counting forward shows by far the flattest curves, although for Hispanic individuals, it seems better at detecting differences in ability, and is noticeably harder for this group (given the curve is shifted so far to the right). We see similar, but more subtle results for counting backwards. The "trails" item seems to perform relatively well for all groups, but non-Hispanic Black and Hispanic individuals show notably steeper curves compared to non-Hispanic Whites. For correctly identifying a rhinoceros, the group that seems to have the easiest time with this is Hispanics, while

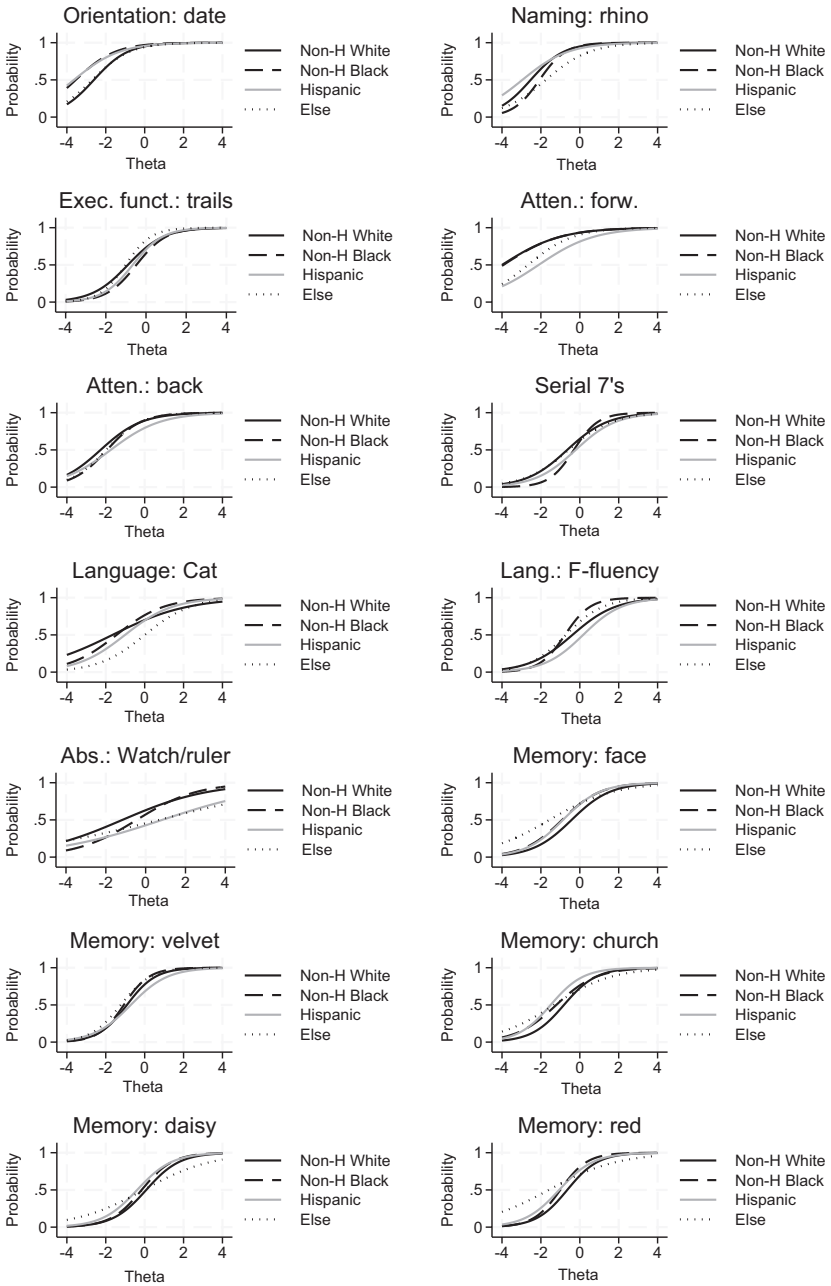

Figure 2. Step 3 part 2; item response curves for items that show different thresholds/loadings across racial group (NB: "theta" is latent test-taking ability).

Table 5. Long and Short Form of MoCA

Domain	Full	Short
Orientation	Month	Month
	Date	
Naming	Rhino	
Visuospatial	Contour	Contour
	Numbers	Numbers
	Hands	Hands
Executive function	Trails	
Memory	Face	
	Velvet	
	Church	
	Daisy	
	Red	
Attention	Forward	
	Backward	
	Serial7	
Language	Fluency	
	Cat	
Abstraction	Watch/Ruler	
Range	0 to 30	0 to 4
Mean	23.06	3.30
SD	4.37	0.80
Mean, by race		
White	23.36	3.30
Black	19.51	2.91
Hispanic	18.66	2.85
Else	22.36	3.25
Mean, by race, standardized		
White	0.06	0.00
Black	−0.82	−0.48
Hispanic	−1.02	−0.56
Else	−0.17	−0.06
R^2 with race/ethnicity	0.15	0.04

Note: MoCA = Montreal Cognitive Assessment.

the "else" and non-Hispanic Black respondents have the most difficult time; for all groups, however, one answers this question successfully even at very low levels of ability. Finally, for knowing the date, this seems to be much harder, and much more discriminating, for White respondents. In short, no single group is always "better" at all tests compared to all others, and the patterns of item response suggest highly variable item behavior by group for these 14 items. In sum, the results show that for all but the four items that we retain, there is considerable variability in item functioning.

In Table 5, we show descriptive statistics for the full and short version of the MoCA-SA. Importantly, observe the *R*-squared value for the association between race and each form of the measure. The *R*-squared for the long form is approximately double that of the short form. This may be due to greater measurement precision in the long form, but it could also be due to a reduction in measurement variance across groups. The two measures are also moderately correlated with one another ($r = 60$), and this is true for non-Hispanic White ($r = 0.60$), non-Hispanic Black ($r = 0.52$), Hispanic ($r = 0.65$), and "other" individuals ($r = 0.62$). The test–retest correlation was considerably lower for the short ($r = 0.38$) versus long version ($r = 0.74$), and this was true for non-Hispanic White ($r = 0.32$ vs $r = 0.66$), non-Hispanic Black ($r = 0.39$ vs $r = 0.75$), Hispanic ($r = 0.38$ vs $r = 0.73$), and "else" groups ($r = 0.24$ vs $r = 0.53$).

Table 6 shows results for consequential outcomes: self-reported dementia, IADL scores, proxy reports of dementia as a contributing cause of death, and NDI-recorded dementia as a cause of death. The short and long forms are always significantly associated with the dependent variable, where higher scores predict a lower probability/level of negative outcomes. Some controls are significant when the full MoCA-SA is included in the model, but not for the short form—and vice versa. For instance, years of education are significantly associated with IADL problems when using the short form, but not the long form, yet is associated with NDI-recorded dementia death when using the full form, but not the short form. Because the coefficients cannot be compared directly across these models, we plotted predicted probabilities for each model as visualized in Figure 3. We can see that for the short form and the long forms, the lines sit virtually on top of one another, indicating the same slope (and thus, association) between cognition and consequential outcomes. However, note that at very low levels of cognition, we see some separation between the short and long form—the long form seems to be slightly better at discriminating between respondents at these low levels, even though the confidence intervals at this level overlap.

Discussion

In the preceding analysis, we found that some items in NSHAP's MoCA-SA function differently across racial and ethnic groups. However, there were items across different cognitive domains that seemed unbiased across groups when capturing cognitive differences. The items that seemed to function well across groups are, in some sense, items that we would expect to reflect common experiences of everyday life in all groups—knowing what month of the year it is (the space of guessing is narrowed down considerably just by knowing the season outside), and knowing how to read an analog clock (though we may not be able to take this cross-cultural validity for granted in more recent generations as digital clocks become more common). It is difficult to explain the behavior of every item for every group, where the items behave differently by group, without the benefit of cognitive interviewing. However, for some, there may be intuitive hypotheses (e.g., the F-fluency question is not as effective at capturing ability among Hispanics, possibly because there are fewer words beginning with "F" in Spanish).

When we created a short form of the MoCA-SA from the measures that function well across groups, we found that the predictive power of the short form was relatively comparable to the long form, except at very low scores, where we observed some separation in predicted probabilities. Furthermore, the association between cognition and race/ethnicity was reduced considerably for the short form, suggesting that some of the differences in cognitive function were due to less-inclusive measures in the longer form of the MoCA-SA. In general, what our findings suggest is that if a researcher is interested in differences by race, or if race and ethnicity are covariates or moderators in analyses featuring cognition, they could carry

Table 6. Predicting Consequential Outcomes With Full and Short Form of MoCA

Cognitive and physical function:	Dementia		IADL		Dementia death—proxy		Dementia death—NDI	
Model:	Firth logit		OLS		Firth logit		Firth logit	
Time:	R3		R3		R2 to R3		R2 to R3	
	Full	Short	Full	Short	Full	Short	Full	Short
MoCA	−1.19***	−0.88***	−0.15***	−0.11***	−1.10***	−0.67***	−1.24***	−0.73***
Race								
White	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Black	−0.71*	−0.18	−0.03	0.03	−0.28	0.28	−1.44**	−0.8
Hispanic	−1.31**	−1.00*	−0.08***	−0.03	−0.01	0.34	−1.03	−0.75
Else	−1.08	−0.69	−0.04	0.00	1.09*	1.31*	0.06	0.15
Years of education	0.08*	−0.02	0.00	−0.01***	0.15***	0.09**	0.12**	0.05
Female	−0.35	−0.39	0.08***	0.06***	0.54*	0.49	0.07	0.05
Age	0.03**	0.06***	0.01***	0.01***	0.06**	0.08***	0.13***	0.15***
Constant	−7.40***	−7.73***	−0.26***	−0.27***	−10.84***	−11.73***	−15.84***	−16.66***
N	4,561	4,561	4,574	4,574	3,330	3,330	3,330	3,330

Notes: IADL = instrumental activities of daily living; MoCA = Montreal Cognitive Assessment; NDI = National Death Index; OLS = ordinary least squares.
*$p < .05$.
**$p < .01$.
***$p < .001$.

out their analyses with both the long and short forms of the MoCA-SA as a robustness check. If the researcher finds that the associations change substantially from one version of the MoCA-SA to the next, it may indicate that the results do not withstand a test of measurement invariance.

The methods here may also provide a template for exploring other kinds of measurement invariance by other groups that the researcher is concerned about (see Piedra et al., 2023). If so inclined, the researcher could also carry out multiple-group structural equation modeling, allowing select items from the full MoCA-SA to vary by group. This retains the entire spectrum of MoCA-SA measures but allows each group to have specific weights and intercepts for specific items. The four items above could thereby function as "anchors" across groups, essentially creating a harmonized score through latent variable modeling (Kobayashi et al., 2021). This approach would also provide added statistical efficiency by estimating latent cognition directly, rather than proxying it using a summative score. It also allows us to retain information from *all* the questions, while allowing for cross-group measurement variance. This approach may therefore be the most scientifically sound, and all else equal, we recommend it above other options. However, it is also the most computationally and technically demanding, and therefore prone to error, as well as unnecessary for the arguments of a single paper. The long-form–short-form robustness check, suggested above, may be a simpler and quicker way to check one's results in many cases—however much it restricts one to a limited pool of items.

Although we aimed to provide useful recommendations above, note that we were limited in terms of what analyses we could do, particularly by subgroup size. Each of the groups considered here is not a monolith, and it is possible that we would see just as much variability in these measures within groups as between. The Hispanic population is heterogeneous in terms of country of origin as well as the modal characteristics of each immigrant wave from a sending country (for instance, whether they primarily arrived as refugees, their average socioeconomic status in the sending country, etc.). The older Black population should also, ideally, be disaggregated. There are cultural differences by region, origin, migration histories, and so on. We also did not have the requisite sample size for considering South Asian, East Asian, or Indigenous older adults, and so can say nothing about any of these populations.

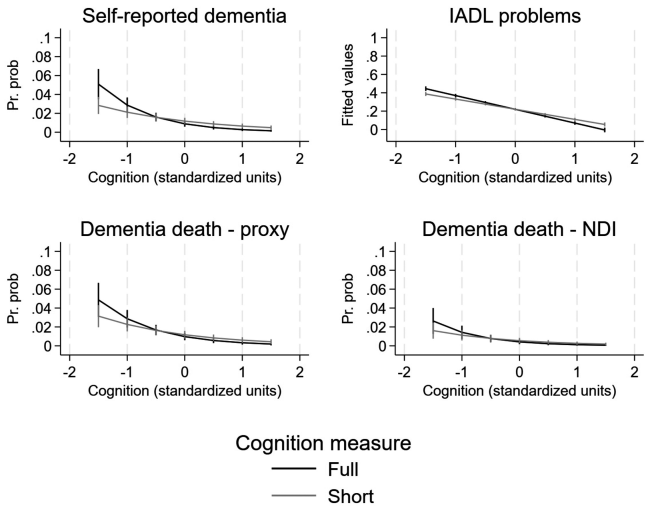

Figure 3. Predicted probabilities for long versus short version of Survey Adjusted version of the Montreal Cognitive Assessment (MoCA-SA). IADL = instrumental activities of daily living; NDI = National Death Index.

Conclusions

Going forward, researchers should not limit ourselves when discussing issues of measurement invariance for variables in NSHAP, to a conversation about the MoCA-SA. The results here are part of a broader discourse about how to construct

good measures that are valid across different groups. We are in a historical moment where inclusion and diversity are central to the advancement of our field, but there is always a danger that what will come of these conversations will be superficial for our science and will not encourage us to go deeper into the refinement, creation, and testing of new measures. This document aims to provide advice to users of the NSHAP data, and accordingly, we cannot go further at this time. But we hope our results will be useful for those who are working to create deeper and more lasting change toward inclusive research on cognitive health and aging.

Funding

The National Social Life, Health, and Aging Project is supported by the National Institute on Aging and the National Institutes of Health (R01AG043538; R01AG048511; R37AG030481). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

None.

References

Adana Díaz, L., Arango, A., Parra, C., Rodríguez-Lorenzana, A., & Yacelga-Ponce, T. (2021). Impact of educational level on versions (basic and complete) of the Montreal Cognitive Assessment. *Dementia and Geriatric Cognitive Disorders*, 50(4), 341–348. https://doi.org/10.1159/000518747

Allison, P. (2012). *Logistic regression for rare events | statistical horizons*. https://statisticalhorizons.com/logistic-regression-for-rare-events/

Alzheimer's Association. (2020). *Race, ethnicity, and Alzheimer's (fact sheet)*. https://aaic.alz.org/downloads2020/2020_Race_and_Ethnicity_Fact_Sheet.pdf

Baumgart, M., Snyder, H. M., Carrillo, M. C., Fazio, S., Kim, H., & Johns, H. (2015). Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimer's & Dementia*, 11(6), 718–726. https://doi.org/10.1016/j.jalz.2015.05.016

Centers for Disease Control. (2016, January 1). *U.S. burden of Alzheimer's disease, related dementias to double by 2060*. Author. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/media/releases/2018/p0920-alzheimers-burden-double-2060

Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation & the Health Professions*, 28(2), 212–232. https://doi.org/10.1177/0163278705275342

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. https://doi.org/10.2307/2336755

Fuller-Thomson, E., & Ahlin, K. M. (2022). A decade of decline in serious cognitive problems among older Americans: A population-based study of 5.4 million respondents. *Journal of Alzheimer's Disease*, 85(1), 141–151. https://doi.org/10.3233/JAD-210561

Global Burden of Disease 2016 Dementia Collaborators. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology*, 18(1), 88–106. https://doi.org/10.1016/S1474-4422(18)30403-4

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *The global burden of disease 2000 in aging populations. Research Paper*.

Kobayashi, L. C., Gross, A. L., Gibbons, L. E., Tommet, D., Sanders, R. E., Choi, S.-E., Mukherjee, S., Glymour, M., Manly, J. J., Berkman, L. F., Crane, P. K., Mungas, D. M., & Jones, R. N. (2021). You say tomato, I say radish: Can brief cognitive assessments in the U.S. Health Retirement Study be harmonized with its international partner studies? *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 76(9), 1767–1776. https://doi.org/10.1093/geronb/gbaa205

Kotwal, A. A., Schumm, L. P., Kern, D. W., McClintock, M. K., Waite, L. J., Shega, J. W., Huisingh-Scheetz, M. J., & Dale, W. (2015). Evaluation of a brief survey instrument for assessing subtle differences in cognitive function among older adults. *Alzheimer Disease and Associated Disorders*, 29(4), 317–324. https://doi.org/10.1097/WAD.0000000000000068

Larson, E. B., Yaffe, K., & Langa, K. M. (2013). New insights into the dementia epidemic. *New England Journal of Medicine*, 369(24), 2275–2277. https://doi.org/10.1056/NEJMp1311405

Lindau, S. T., Schumm, L. P., Laumann, E. O., Levinson, W., O'Muircheartaigh, C. A., & Waite, L. J. (2007). A study of sexuality and health among older adults in the United States. *New England Journal of Medicine*, 357(8), 762–774. https://doi.org/10.1056/NEJMoa067423

Luck, T., Roehr, S., Jessen, F., Villringer, A., Angermeyer, M. C., & Riedel-Heller, S. G. (2015). Mortality in individuals with subjective cognitive decline: Results of the Leipzig Longitudinal Study of the Aged (LEILA75+). *Journal of Alzheimer's disease*, 48(s1), S33–S42. https://doi.org/10.3233/JAD-150090

Manly, J. J., Jones, R. N., Langa, K. M., Ryan, L. H., Levine, D. A., McCammon, R., Heeringa, S. G., & Weir, D. (2022). Estimating the prevalence of dementia and mild cognitive impairment in the US: The 2016 Health and Retirement Study Harmonized Cognitive Assessment Protocol Project. *JAMA Neurology*, 79, 1242. https://doi.org/10.1001/jamaneurol.2022.3543

McSorley, V. E., Bin, Y. S., & Lauderdale, D. S. (2019). Associations of sleep characteristics with cognitive function and decline among older adults. *American Journal of Epidemiology*, 188(6), 1066–1075. https://doi.org/10.1093/aje/kwz037

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nasreddine, Z. S., Phillips, N. A., Bédirian, 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 Geriatrics Society*, 53(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *Patient*, 7(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

O'Driscoll, C., & Shaikh, M. (2017). Cross-cultural applicability of the Montreal Cognitive assessment (MoCA): A systematic review. *Journal of Alzheimer's Disease*, 58(3), 789–801. https://doi.org/10.3233/JAD-161042

O'Muircheartaigh, C., Eckman, S., & Smith, S. (2009). Statistical design and estimation for the National Social Life, Health, and Aging Project. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 64B(suppl_1), i12–i19. https://doi.org/10.1093/geronb/gbp045

O'Muircheartaigh, C., English, N., Pedlow, S., & Kwok, P. K. (2014). Sample design, sample augmentation, and estimation for Wave 2 of the NSHAP. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 69(Suppl 2), S15–S26. https://doi.org/10.1093/geronb/gbu053

Parker, C., & Philp, I. (2004). Screening for cognitive impairment among older people in Black and minority ethnic groups. *Age and Ageing*, 33(5), 447–452. https://doi.org/10.1093/ageing/afh135

Perry, B. L., McConnell, W. R., Peng, S., Roth, A. R., Coleman, M., Manchella, M., Roessler, M., Francis, H., Sheean, H., & Apostolova, L. A. (2022). Social networks and cognitive function: An

evaluation of social bridging and bonding mechanisms. *Gerontologist*, 62(6), 865–875. https://doi.org/10.1093/geront/gnab112

Piedra LM, Iveniuk J, Howe MJK, Pudelek K, Marquez DX. (2024). Addressing cognitive assessment disparities among Hispanic adults: Adapting the MoCA-SA for improved accuracy and accessibility among Spanish-speakers. *The Journals of Gerontology: Series B*, gbae036. https://doi.org/10.1093/geronb/gbae036

Rajan, K. B., Weuve, J., Barnes, L. L., McAninch, E. A., Wilson, R. S., & Evans, D. A. (2021). Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060). *Alzheimer's & Dementia*, 17(12), 1966–1975. https://doi.org/10.1002/alz.12362

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). The problem of bias in psychological assessment. In C. R. Reynolds, R. A. Altmann, & D. N. Allen (Eds.), *Mastering modern psychological testing: Theory and methods* (pp. 573–613). Springer International Publishing. https://doi.org/10.1007/978-3-030-59455-8_15

Shega, J. W., Sunkara, P. D., Kotwal, A., Kern, D. W., Henning, S. L., McClintock, M. K., Schumm, P., Waite, L. J., & Dale, W. (2014). Measuring cognition: The Chicago cognitive function measure in the National Social Life, Health and Aging Project, wave 2. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 69(Suppl_2), S166–S176. https://doi.org/10.1093/geronb/gbu106

Shuttleworth-Edwards, A. B. (2016). Generally representative is representative of none: Commentary on the pitfalls of IQ test standardization in multicultural settings. *Clinical Neuropsychologist*, 30(7), 975–998. https://doi.org/10.1080/13854046.2016.1204011

StataCorp. (2021). *Stata statistical software: Release 17* [Computer software]. StataCorp LLC.

Tilvis, R. S., Kähönen-Väre, M. H., Jolkkonen, J., Valvanne, J., Pitkala, K. H., & Strandberg, T. E. (2004). Predictors of cognitive decline and mortality of aged people over a 10-year period. *Journals of Gerontology, Series A: Biological Sciences and Medical Sciences*, 59(3), M268–M274. https://doi.org/10.1093/gerona/59.3.m268

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 54(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Waite, L. J., Hawkley, L., Kotwal, A. A., O'Muircheartaigh, C., Schumm, L. P., & Wroblewski, K. (2021). Analyzing birth cohorts with the National Social Life, Health, and Aging Project. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 76(Suppl_3), S226–S237. https://doi.org/10.1093/geronb/gbab172

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education*, 80(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Willis, G. (2005). *Cognitive interviewing*. Sage Publications. https://doi.org/10.4135/9781412983655