Research Article

Psychometric reliability, validity, and generalizability of 3MSE scores among American Indian adults: the Strong Heart Study

Astrid M. Suchy-Dicey[1,2] , Thao T. Vo[3], Kyra Oziel[2], Dedra S. Buchwald[2], Lonnie A. Nelson[4], Steven P. Verney[5] and Brian F. French[3]

[1]Huntington Medical Research Institutes, Pasadena, CA, USA, [2]Washington State University, Elson S Floyd College of Medicine, Seattle, WA, USA, [3]Washington State University, College of Education, Pullman, WA, USA, [4]Washington State University, College of Nursing, Spokane, WA, USA and [5]University of New Mexico, Albuquerque, NM, USA

Abstract

Objective: Modified Mini-Mental State Examination (3MSE) is often used to screen for dementia, but little is known about psychometric validity in American Indians. **Methods:** We recruited 818 American Indians aged 65–95 for 3MSE examinations in 2010–2013; 403 returned for a repeat examination in 2017–2019. Analyses included standard psychometrics inferences for interpretation, generalizability, and extrapolation: factor analysis; internal consistency-reliability; test-retest score stability; multiple indicator multiple cause structural equation models. **Results:** This cohort was mean age 73, majority female, mean 12 years education, and majority bilingual. The 4-factor and 2nd-order models fit best, with subfactors for orientation and visuo-construction (OVC), language and executive functioning (LEF), psychomotor and working memory (PMWM), verbal and episodic memory (VEM). Factor structure was supported for both research and clinical interpretation, and factor loadings were moderate to high. Scores were generally consistent over mean 7 years. Younger participants performed better in overall scores, but not in individual factors. Males performed better on OVC and LEF, females better on PMWM. Those with more education performed better on LEF and worse on OVC; the converse was true for bilinguals. All differences were significant, but small. **Conclusion:** These findings support use of 3MSE for individual interpretation in clinic and research among American Indians, with moderate consistency, stability, reliability over time. Observed extrapolations across age, sex, education, and bilingual groups suggest some important contextual differences may exist.

Keywords: American Indian; health disparities; neuropsychology; cognition; psychometrics; 3MSE

(Received 4 August 2023; final revision 14 November 2023; accepted 11 December 2023; First Published online 24 January 2024)

Introduction

The Mini-Mental State Examination (MMSE) is often used to screen for dementia (Tsoi et al., 2015). With tens of thousands of scientific publications in the 40 years since its inception, MMSE has become one of the most widely used cognitive tests globally (Folstein et al., 1983). The Modified Mini-Mental State Examination (3MSE, aka 3MS) is an adaptation of MMSE, with four additional test items (long-term memory, semantic fluency, abstract thinking, and one additional verbal memory task), standardized administration and scoring instructions, and an expanded, more interpretable score range (McDowell et al., 1997; Teng & Chang Chui, 1987; Tombaugh et al., 1996). Many psychometric studies of MMSE or 3MSE have focused on single populations, with an inclusion bias towards Non-Hispanic Whites (NHW), although some normative score differences comparing racial/ethnic groups have been published (Rapp et al., 2003; Ryan et al., 2019; Sink et al., 2015). Prior research in American Indians has demonstrated population-wide performance on 3MSE and other cognitive tests is overall low, compared with NHW and other groups, suggesting potential test invalidity for minoritized groups (A. Suchy-Dicey et al., 2022; A. Suchy-Dicey et al., 2020; Verney et al., 2019). Socioeconomic disparities account for some of the observed discrepancies in race-associated cognitive performance (Galvin et al., 2021; McCann et al., 2018; Zahodne et al., 2021; Zahodne et al., 2019), indicating race may be a proxy for underlying social and institutional inequities (Avila et al., 2020). However, incomplete social determinant models maintain race as a convenient, albeit inadequate, substitute for complex, latent constructs driving differences in test performance.

Normative adjustment or standardization for socio-demographic factors – including race – continues to be a common contextualization tool when assessing cognition in diverse populations (Lezak, 2012). Nevertheless, normalization carries key limitations that are easily overlooked, and neuropsychologists are divided on the best methods for comprehensive and equitable cognitive evaluation of minoritized elders (Brickman et al., 2006;

Byrd & Rivera-Mindt, 2022; Fernandes et al., 2022). Furthermore, racial and ethnic representation in neuropsychology research has been critically inadequate. For example, despite heavy lifetime risk for vascular and Alzheimer's dementias (Kim et al., 2012; Livingston et al., 2020), population-based psychometrics and normative score estimates have yet to be undertaken among any U.S. Indigenous populations, including American Indians, for most standardized cognitive tests, including 3MSE (Mehta & Yeo, 2017; Verney et al., 2019).

Validity of MMSE and 3MSE scores

Psychometric evaluation of validity is critical to the interpretation of test score distributions as well as for making individual clinical decisions, including claims about cognitive domains or functional status. An integrated method to examine multiple points of validity is the Kane framework (Cook et al., 2015; Kane, 2013), wherein inference for valid interpretation involves evaluation of model fit, including structural components such as sub-scores; generalizability inference covers consistency, reliability, stability; and inference of extrapolation covers differences due to population stratification, for example by sociodemographics. Overall, this framework provides a more integrated, structured view of psychometric validity. Existing points of evidence in this validity framework for 3MSE in the majority population include inter-rater reliability (Pearson's $r = 0.98$, (Bassuk & Murphy, 2003; Teng & Chang Chui, 1987)), internal consistency-reliability (Cronbach's alpha = 0.82 to 0.87, (McDowell et al., 1997; Tombaugh et al., 1996)), test-retest reliability or score stability ($r = 0.80$, (Grace et al., 1995)), and inter-test comparisons (MMSE: MCI threshold < 28/30, ROC 0.74; 3MSE: MCI threshold < 95/100, ROC 0.85 (Van Patten et al., 2019)). Compared to MMSE, 3MSE explains more variance in overall cognition and domain-specific tasks related to language and memory but not for attention, visuospatial, or executive function (Van Patten et al., 2019).

Cultural factors in MMSE and 3MSE testing

Intended as a dementia screening and evaluation tool, MMSE and 3MSE sample multiple cognitive domains likely to reflect a range of human experiences common among persons from diverse backgrounds (Teng & Chang Chui, 1987). However, the dearth of formal investigations and adaptations among cultural minorities raises uncertainty about its cross-cultural applications. Differential scores across groups of Canadian native language speakers suggest key, latent linguistic and cultural influences on test performance (Bravo & Hébert, 1997). Furthermore, translation concerns have also been raised, in part because of modifying effects of education on bilingualism in test scores (McDowell et al., 1997). Education is a critical contextual factor in cognitive test performance and score interpretation (Bassuk & Murphy, 2003; Bravo & Hébert, 1997; Tschanz et al., 2002), and is also likely to impact score performance outside of any effect on language. Differences in educational tenure and quality can impact cognitive status and test performance via test familiarity and anxiety. Overall, both linguistic and educational differences appear to influence MMSE and 3MSE test contextualization and adaptation in minoritized populations, with the exclusion of one or the other factor likely to result in information bias and misidentification of case status (Manly et al., 1998). Thus, to enable valid interpretations for MMSE or 3MSE scores, key educational, linguistic, and sociodemographic factors must be considered.

MMSE and 3MSE testing in American Indian elders

Although normative score ranges in NHW, African American, and Hispanic/Latino populations have been published (Rapp et al., 2003; Ryan et al., 2019; Sink et al., 2015), there is limited prior information on MMSE or 3MSE scores among American Indian populations. A small study of MMSE in 140 American Indian elders described summary scores, with cross-sectional associations to sex and education but not age (L. L. Jervis et al., 2007), as well as longitudinal determinant associations from age, education, income, and healthcare access (L. L. Jervis et al., 2010). In that population-based study, MMSE scores were generally low, compared with published scores for other populations, but not consistent with the approximate 11% who had self-identified cognitive impairment, suggesting standard clinical thresholds may need reconsideration (Crum et al., 1993). The Women's Health Initiative also published 3MSE score ranges for $N = 26$ American Indian or Alaska Native females (age 65–80) out of 7480 (0.3%) total participants (Rapp et al., 2003), a critical underrepresentation for a group numbering an estimated 10 M and comprising >3% of the US population (Census, 2020). In that study, 3MSE score ranges for American Indians and Alaska Natives were not substantively different from other racial/ethnic groups and did not differ substantively by age group. However, in a much larger, population-based examination from the Strong Heart Study (SHS), with $N = 818$ American Indian participants from both sexes (aged 65–95), 3MSE score ranges were overall lower than previously published scores, including those of the same-age in the general population as well as those of American Indian and Alaska Native identity in the Women's Health Initiative; in the SHS report, 3MSE mean scores fall below standard clinical thresholds for mild cognitive impairment and dementia (Ryan et al., 2019; Tombaugh et al., 1996; Verney et al., 2019).

Further examination of score ranges across SHS population strata detected cross-sectional associations of test performance with age and education but not sex or language use (bilingual). Of key consideration, a majority of living American Indian elders in this age range are survivors of Indian boarding schools, one component of a 300-year U.S. federal policy which resulted in considerable childhood mortality, physical and psychological trauma, cultural inculcation, linguistic alienation, cultural erasure, and poor quality learning experiences (Brave Heart & DeBruyn, 1998; Verney et al., 2015). Accordingly, educational tenure is likely to be a poor representation of true crystallized intelligence and intellectual achievement in this population (A. Suchy-Dicey et al., 2022), and many affected elders are likely to have significant test anxiety and other performance considerations. These findings emphasize the importance of independent characterization of psychometric validity and performance factors in cognitive testing in this population.

Theory

This report comprises several psychometric inferences of 3MSE validity in a large, heterogeneous, population-based cohort of American Indian participants aged 65–95 years (Kane, 2013; A. M. Suchy-Dicey et al., 2016). First, a scoring inference was assessed by examining 3MSE factor structure to evaluate how responses on the instrument match the underlying theory on which MMSE and 3MSE were constructed. Second, a generalizability inference was examined for internal consistency-reliability, and for test-retest score stability in 3MSE scores. Third, an extrapolation inference examined group differences and differential indicator functioning

to evaluate differences in individual tasks across age, sex, education, and bilingual strata. Altogether, this work provides insights into how well 3MSE measures cognitive functioning overall and within specific domains; whether cultural, linguistic, or other test adaptations need to be considered; and whether any cautions or recommendations for clinical interpretation are warranted.

Methods

Setting

In 1989–1991, the Strong Heart Study recruited middle-aged individuals claiming heritage from any of 13 partnering American Indian tribes and communities in the Northern Plains, Southern Plains, and Southwest U.S. (Lee et al., 1990). From 2010 to 2013 and again in 2017 to 2019, all surviving Strong Heart Study participants (then aged 65–95) were invited to undergo cognitive testing, brain imaging, and clinical examination as part of the Cerebrovascular Disease and its Consequences in American Indians (CDCAI) study (A. M. Suchy-Dicey et al., 2022; A. M. Suchy-Dicey et al., 2016; A. M. Suchy-Dicey et al., 2017). Every effort was made for complete cohort ascertainment at all phases, with 75–86% successful recruitment rates. In accordance with the Helsinki declaration and with sovereign tribal agreements, all participating tribal review boards, Indian Health Service, and institutional IRBs approved study activities, and all participants provided written informed consent.

Measures

Both CDCAI examination visits (Visit 1, Visit 2) involved similar protocols. Participants completed self-reported history and behavioral questionnaires with measures including sex (male, female), age (years), years of education (12 or fewer; 13 or more), and speaking ability in their Native language (moderately or very well; not at all or a little). Combined with the cohort eligibility requirement that all participants be fluent in English, the language variable represents bilingual or multi-lingual speaking ability. Field center staff trained in administration of neuropsychological examinations conducted cognitive testing interviews, including 3MSE.

The 3MSE consists of 40 individual tasks scored on a summary 100-point scale (Teng & Chang Chui, 1987). The overall score is intended to represent overall or general cognition, with subitems for attention, concentration, long and short-delay verbal memory, orientation to time and place, expressive and receptive language, semantic fluency, constructional praxis, and abstract thinking. In this test, participants were asked to state the place and date of their birth, including town, state, year, month, and day (orientation: 5 points). Registration trials for a verbal memory task involved repeating three words with up to five repetitions for encoding, although only the first trial was scored (registration: 3 points). The mental reversal tasks involved asking participants to count backward from 5, and to spell a five-letter word backward (reversal: 7 points). The first (short delay) recall trial asked participants to repeat the three memory-task words, with semantic and phonemic cues provided as needed (recall1: 9 points). Temporal orientation asked for year, month, date, day of the week, and season (temporal: 15 points). Spatial orientation asked for state, county, city, and specific location/building (spatial: 5 points). Participants were asked to name five body parts, identified by pointing (name: 5 points). A timed task asked

participants to name as many four-legged animals as possible in thirty seconds (animals: 10 points). A similarities or abstract thinking task asked participants to provide an expression for how three sets of two items are similar (similarities: 6 points). A verbal repetition task involved repeating a phrase (repetition: 5 points). A written instructional task involved obeying instructions to "close your eyes" (obey: 3 points). A writing task involved writing a five-word sentence within one minute (write: 5 points) and a simple figure was provided for copying within one minute (pentagons: 10 points). A verbal instruction task was given, wherein participants were asked to grasp, fold, and return a sheet of paper (command: 3 points). Finally, the second (long delay) recall trial of the three memory-task words, with cues provided if needed, was administered (recall2: 9 points). Two trained reviewers evaluated item performance and individual task scores for accuracy and consistency over the entire cohort. A panel of neuropsychology and epidemiology experts reviewed any items with scoring inconsistency, disagreement, or uncertainty to reach consensus scoring for those items.

Analysis: Inference of validity in score interpretation

Inference of scoring evaluated whether summary (unifactorial model) or multiple scores (multifactorial models) accurately capture intended test constructs, using confirmatory factor analysis with robust maximum likelihood estimation (Finney & DiStefano, 2013). Model fit was evaluated using standard indices and criteria, including Root Mean Square Error of Approximation (RMSEA < 0.08) (Brown, 2015), Standardized Root Mean Square Residual (SRMR < 0.05) (Muthén, 1989), and Comparative Fit Index (CFI > 0.90) (Browne & Cudeck, 1992). The chi-square test, which is sensitive to sample size, assessed model-data fit with a significant value ($p < 0.01$), indicating a lack of fit. First, a unidimensional (1-factor) model was examined, based on MMSE/3MSE creator intent and manual instructions (Teng & Chang Chui, 1987). Additionally, 3-and 4-factor models were considered based on secondary findings from other populations. The 3-factor model included factors for learning, recall, and executive functioning (LREF: recall 1, recall 2, animal, repetition, similarities, reversal, registration, temporal tasks); orientation (OA: name, orientation, spatial tasks); and psychomotor ability (PA: command, obey, write, pentagons tasks) (Cappeliez et al., 1996; Klyce et al., 2021). The 4-factor model included verbal-episodic memory (VEM: recall 1, recall 2); language and executive function (LEF: animal, command, name, obey, repetition, similarities tasks); psychomotor and working memory (PWM: reversal, write tasks); and orientation and visuo-construction (OVC: orientation, pentagons, registration, spatial, temporal tasks) (Ryan et al., 2019). As an extension of the 4-factor model, a second-order model was examined (Petrill et al., 1996).

Analysis: Inference of generalizability and reliability

Inference of generalizability assesses whether test scores are a legitimate representation of the assessment universe, in accordance with the larger validity framework (Kane, 2013). For inference of generalizability, we used omega (ω) coefficient as an estimate of internal consistency-reliability to quantify the amount of random measurement error that exists in a given score (McDonald, 1999). Omega was used as the reliability estimate because it is more robust than Cronbach's alpha coefficient against potentially unmet assumptions, such as that all test items have the same relationship with the underlying construct, also called tau equivalence. Omega
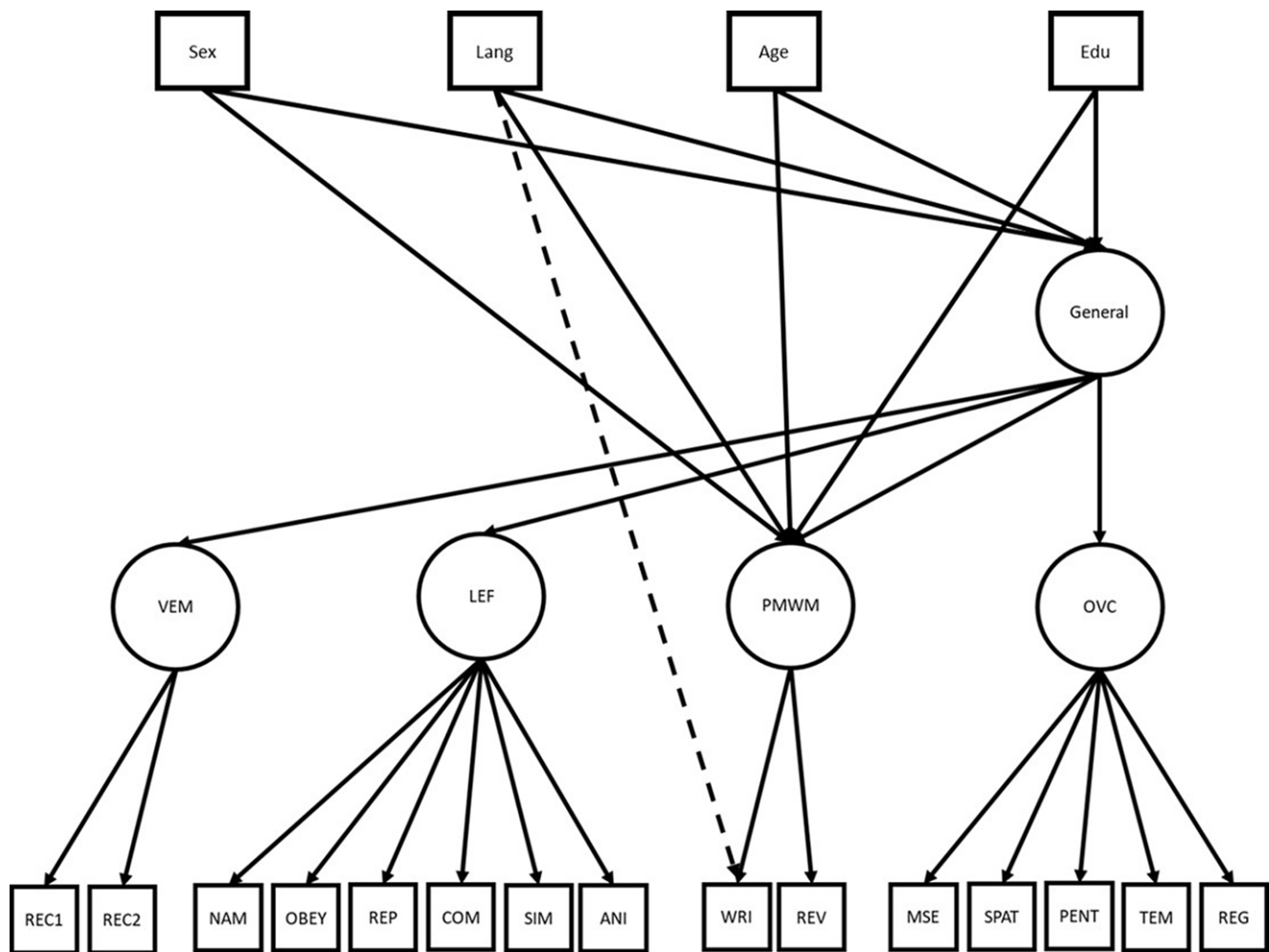
Figure 1. MIMIC Model for detecting subgroup differences in Modified Mini Mental Status Examination (3MSE) scores, with an example given for the tasks "Write" while controlling for Overall (total score) aka "General" and for four first-order factors. Lang represents bilingual speaking ability; Edu: tenure of formal education in years; VEM: Verbal Episodic Memory; LEF: Language and Executive Function; PMWM: Psychomotor & Working Memory; OVC: Orientation & Visuo-construction. Individual items represent single tasks on 3MSE, including Rec1: short-delay word list, Rec2: long delay word list, Nam: name body parts, Obey: written instructional task, Rep: verbal repetition, Com: verbal instructional task, Sim: abstract similarities, Ani: animal naming, Wri: writing task, Rev: mental reversal, Mse: orientation and personal details, Spat: spatial orientation, Pent: figure drawing, Tem: temporal orientation, Reg: registration.

values >0.9 and >0.8 are considered adequate for valid interpretation in clinical and research evaluation purposes, respectively (Nunnally, 1994).

Lambda (λ) coefficients, or factor loadings, were estimated to reflect the correlation between an individual item, i.e., a single task on the 3MSE, and its given factor, e.g., one of the four groups of tasks (VEM: verbal-episodic memory, etc.). Lambda ranges from 0 to 1, with higher values representing a greater correlation between an individual item and the overall factor; values of lambda >0.3 are conventionally considered to reflect a moderate degree of correlation (Comrey & Lee, 1992; Pett et al., 2003).

Similarly, test-retest reliability or score stability was measured using Pearson's correlation coefficient (r) to examine whether conditions of observation, or occasion of administration, unduly influence variance across subgroups or, to put it another way, represent samples of the universe of conditions for the domains assessed. Coefficients were estimated across strata for sex, age, language, education, and by examination visit; coefficients above 0.60 are consistent with score stability (Diedenhofen & Musch, 2015). The mean 6.7-year interval in this study is similarly

structured to prior psychometric, stability studies on 3MSE with multi-year intervals (Tombaugh, 2005).

Analysis: Inference of extrapolation

Inference of extrapolation examined whether test scores differ across age, sex, language, and education strata using multiple indicator multiple cause (MIMIC; Figure 1) structural equation models. MIMIC is an extension of confirmatory factor analysis with covariates, such as sociodemographics, and is typically used to contextualize latent variable effects on factors of interest by estimating manifest variable effects. The available (e.g., sociodemographic) variables are named "cause" variables in the model name; however, there is no implication of causal effect, so they are sometimes also called exogenous variables. In step 1, a MIMIC regression model is fitted, with each of the four factors included (age, sex: males vs. females, education: ≤12 or 13+ years, and language/bilingual ability: not at all/a little vs. moderately/ very well) to examine group differences in 3MSE scores. Results of this step in MIMIC modeling include the estimation of regression

Table 1. Select sociodemographic and cognitive test score characteristics for American Indian participants from CDCAI Visit 1 (2010–2013) and Visit 2 (2017–2019)

	Visit 1 $N = 814$	Visit 2 $N = 402$
Age, years	73.0 (5.9)	78.1 (4.7)
64–69.9 years: %	34.2	–
70–74.9 years: %	31.4	32.3
75–79.9 years: %	19.6	37.9
80+ years: %	14.8	29.8
Female: %	67.5	70.1
Education, years	12.6 (2.7)	13.0 (2.6)
≤12 years: %	53.8	49.2
>13 years: %	46.0	50.7
Bilingual / multiple language speaking ability (moderately or very well): %	39.5	34.0
Modified Mini Mental Status Examination (3MSE), overall score	88.4 (9.7)	87.1 (9.5)
Item score		
Orientation	5.0 (0.2)	5.0 (0.2)
Registration	3.0 (0.1)	3.0 (0.2)
Reversal	6.4 (1.2)	6.3 (1.3)
Recall 1	7.6 (1.8)	6.7 (1.7)
Temporal	14.3 (2.2)	14.3 (2.1)
Spatial	4.9 (0.5)	4.8 (0.4)
Name	4.9 (0.4)	4.8 (0.6)
Animal	8.2 (2.2)	8.4 (2.0)
Similarities	2.6 (2.0)	3.2 (2.1)
Repetition	4.7 (0.6)	4.6 (0.9)
Obey	2.9 (0.4)	2.9 (0.4)
Write	4.8 (.7)	4.5 (1.0)
Pentagons	9.0 (1.5)	9.1 (1.6)
Command	2.9 (0.4)	3.0 (0.2)
Recall 2	7.4 (2.1)	6.6 (2.5)

Note: Scores presented as mean (standard deviation) unless otherwise noted. Time between visit 1 and 2 ranged between 7 and 8 years, therefore, there were no participants aged 60–69 at visit 2.

coefficients (γ) or the association of the manifest variable on the latent variable. Group differences are estimated across strata using Cohen's *d* statistic, wherein 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes, respectively (Lenhard & Lenhard,

2016). Benjamini-Hochberg false discovery rate is used to control Type I Error due to multiple comparisons (Benjamini & Hochberg, 1995). In step 2, differences between bilingual and non-bilingual groups, independent of effects from sex, age, and education, were estimated based on linear regression coefficients (β) of the overall score and each factor sub-score, thus estimating a direct, independent path from language factor to each individual item.

All analyses were conducted using M*plus* 2017 (Los Angeles, CA) and SPSS v27 (IBM Corporation, Armonk, NY). Missing or incomplete data affected <1% of cohort participant records, and only complete records were included in analyses. Data from Visit 1 were used for analyses on score interpretation, generalizability, consistency, and stability; data from both Visit 1 and Visit 2 were used for analyses on test-retest score stability, and extrapolation.

Results

Participant characteristics are presented in Table 1. As previously reported, CDCAI recruited $N = 818$ participants in 2010–2013, then aged 64-95, with 86% recruitment from prior SHS visits (A. M. Suchy-Dicey et al., 2016); in 2017–2019, $N = 403$ of these participants, then aged 71-93, returned for a follow-up, representing 78% recruitment (A. Suchy-Dicey et al., 2022). The main causes of attrition between examination visits included mortality and frailty. Very few were missing needed data for these analyses: $n = 4$ (0.5%) at Visit 1, $n = 1$ (0.2%) at Visit 2. The majority were female, had approximately 12 years of education, and were at least minimally bilingual, speaking both English fluently (a study requirement) and their Native language either moderately or very well. Scores on the 3MSE total, subscale, and factor item were similar between the two visits.

The full range of scores on 3MSE summary scores are presented in Figure 2. The distributions at both Visit 1 and Visit 2 overlap considerably. Of note, although the distribution at both Visit 1 and Visit 2 are both highly left-skewed, as is characteristic of this test, few participants achieved maximum scores, reducing the
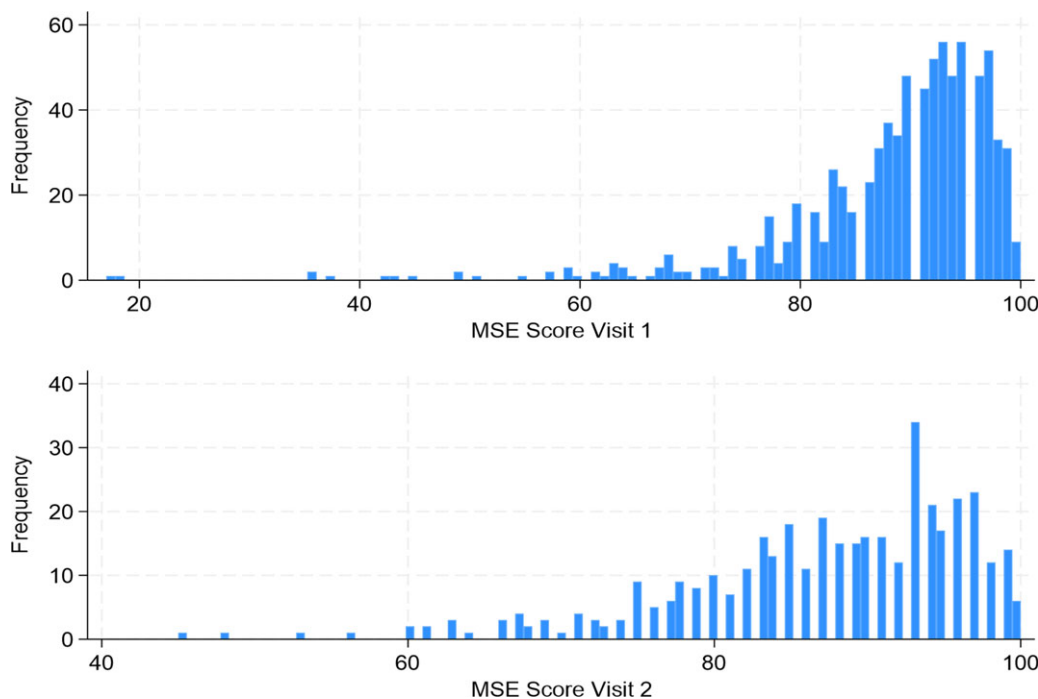
Figure 2. Histogram distributions of Modified Mini Mental Status Examination Scores in American Indian adults (age 65–95) participants of the Strong Heart Study at cognitive examination Visit 1 (2010–2013) and Visit 2 (2017–2019).

Table 2. Model fit indices from exploratory factor analysis of Modified Mini Mental Status Examination (3MSE) scores among American Indian adults, ages 65–95

Model	χ^2	df	CFI	SRMR	RMSEA (CI)
Unidimensional (1-factor)	471.19*	93	0.65	0.15	0.07 (0.06, 0.08)
3-factor	235.28*	87	0.86	0.05	0.05 (0.04, 0.05)
4-factor	179.42*	84	0.91	0.04	0.06 (0.03, 0.04)
Second Order	186.05*	86	0.91	0.04	0.04 (0.03, 0.04)

Note: * = $p < 0.01$. CFI: Comparative Fit Index; SRMR: Square Root Mean Residual; RMSEA (CI): Root Mean Square Estimation Approximation with 95% confidence interval. Note that the second-order model contains the four lower-order factors as well as the primary higher-order factor. Comparison of the nested 4-factor and second-order model χ^2 (2) = 7.08, p-value > 0.05.

Table 3. Standardized pattern coefficients (λ) and internal consistency-reliability estimates (ω) from confirmatory factor analysis of the second-order model of Modified Mini Mental Status Examination (3MSE) scores among American Indian adults, ages 65–95

Second-order model factor indicator	Sub-factor indicator	ω	λ (SE)
Verbal Episodic Memory (VEM)		0.81	
	Recall 1[LREF]		0.81 (0.03)
	Recall 2[LREF]		0.82 (0.03)
Language and Executive Function (LEF)		0.66	
	Animal[LREF]		0.48 (0.04)
	Command[PA]		0.40 (0.09)
	Name[OA]		0.52 (0.07)
	Obey[PA]		0.58 (0.09)
	Repetition[LREF]		0.55 (0.06)
	Similarities[LREF]		0.45 (0.04)
Psychomotor & Working Memory (PMWM)		0.49	
	Reversal[LREF]		0.49 (0.06)
	Write[PA]		0.65 (0.08)
Orientation & Visuo-construction (OVC)		0.54	
	Orientation[OA]		0.34 (0.10)
	Pentagons[PA]		0.46 (0.08)
	Registration[LREF]		0.11 (0.10)
	Spatial[OA]		0.54 (0.09)
	Temporal[LREF]		0.69 (0.06)
Overall (Total score)		0.94	
	VEM		0.65 (0.06)
	LEF		0.96 (0.06)
	PMWM		0.98 (0.07)
	OVC		0.91 (0.07)

Note: SE = Standard Error. Cross-indexing to 3-factor model item groups noted: LREF = Learning, Recall, and Executive Functioning (recall 1, recall 2, animals, repetition, similarities, reversal, registration, temporal); PA = Psychomotor Ability (command, obey, write, pentagons); OA = Orientation and Awareness (name, orientation, spatial).

Table 4. Modified Mini Mental Status Examination (3MSE) scores, with test-retest score stability (visit-to-visit) correlation coefficients (Pearson's r) and cross-group comparisons among American Indian adults, ages 65–95

Group	M (SD) Visit 1	M (SD) Visit 2	Pearson's r
Total	91.2 (6.6)	87.1 (9.5)	0.66
Sex			
Male	91.1 (6.4)	86.6 (8.1)	0.60
Female	91.3 (6.7)	87.4 (10.0)	0.68
Age in years			
64–69.9	91.1 (5.3)	89.6 (7.9)	0.62
70–74.9	91.2.3 (7.6)	86.8 (9.5)	0.67
75–74.9	89.3(7.8)	83.6(10.2)	0.70
80+	90.7 (5.6)	81.5 (11.1)	0.70
Education in years			
≤12 years	88.6 (7.4)	84.1 (10.4)	0.69
>13 years	93.7 (4.7)	89.9 (7.5)	0.47
Native language spoken well			
Not at all	92.2 (6.0)	88.4 (8.5)	0.57
A little but not very well	91.5 (6.8)	87.4 (9.7)	0.69
Moderately well	91.0 (7.4)	85.2 (10.7)	0.72
Very well	89.3 (10.6)	85.8 (9.8)	0.67

Note: M: Mean, SD: standard deviation. All Pearson's correlations or test-retest reliability coefficients were significantly different from 0, $p < 0.05$; all pairwise comparisons of stability coefficients for within-group comparisons were not significant, $p > 0.05$.

parsimonious, and supports the more interpretable total score, this model is used in subsequent analyses.

The lambda (λ) coefficients, or factor loading estimates, for each of the individual 3MSE tasks suggest all except one task (Registration, $\lambda = 0.11$) were >0.3, the conventional threshold for at least moderate correlation between the individual item and its factor. The VEM factor had overall high loadings (>0.8); LEF and PMWM moderate loadings (range 0.4–0.65); and OVC low to moderate with high loading variability (range 0.11–0.69). The Overall factor, representing the second-order for the total score, had reasonably high loadings, ranging from 0.65 to 0.98.

Inference for generalizability: Internal consistency and test-retest score stability

Evaluation of the 4-factor model included the following factors: Verbal Episodic Memory (VEM), Language and Executive Function (LEF), Psychomotor & Working Memory (PMWM), Orientation & Visuo-construction (OVC); the second-order factor was the Overall factor, reflecting the total score (Table 3). The omega (ω) coefficient, or internal consistency-reliability estimate, for the Overall factor (total score), was >0.9, the threshold used for valid interpretation in clinical contexts. The only omega coefficient for a sub-factor >0.8 was VEM, consistent with valid interpretation for research applications. The other omega coefficients (LEF, PMWM, OVC) ranged from 0.49 to 0.66.

Total 3MSE scores were moderate to highly stable between Visit 1 and Visit 2, with Pearson's reliability-stability coefficients (r) ranging from 0.47 to 0.78 (Table 4). The lowest coefficient ($r = 0.47$) was identified among participants with >13 years of education, and the highest ($r = 0.78$) among the oldest group, aged >80 years. All Pearson's reliability-stability coefficients were statistically different from 0 ($p < 0.05$), suggesting that 3MSE scores were not dissimilar – i.e., were stable – between Visit 1 and Visit 2 for all subgroups evaluated. Furthermore, all within-group comparisons were not significant ($p > 0.05$). However, although not significant, males, younger persons, those with more education, and those with lower bilingual ability had poorer reliability-stability coefficients than their counterparts.

likelihood of ceiling effect influencing analyses: $n = 9$ at Visit 1, $n = 6$ at Visit 2, $n = 3$ at both visits.

Inference for score validity: factor structure

Based on standard model fit criteria (CFI >0.9, SRMR <0.05, and RMSEA <0.08), the unidimensional or single-factor model (CFI = 0.65, SRMR = 0.15, RMSEA 95% CI: 0.06–0.08) and the 3-factor model (CFI = 0.86, SRMR = 0.05, RMSEA 95% CI: 0.04–0.05) did not meet fit criteria (Table 2). On the other hand, the 4-factor model (CFI = 0.91, SRMR = 0.04, RMSEA 95% CI: 0.03–0.04) and the second-order model (CFI = 0.91, SRMR = 0.04, RMSEA 95% CI: 0.03–0.04) both met all three model fit criteria. Furthermore, the difference between the 4-factor and second-order models was not significant (χ^2 (2) = 7.08; $p > 0.05$), indicating the model fit is not worse for the more restrictive, second-order model. Because the second-order model has a comparable fit, is more

Table 5. Multiple Indicator Multiple Cause (MIMIC) structural equation model coefficients (γ) and differential item functioning effect size estimates (Cohen's d) for a second-order, 4-factor structure of Modified Mini Mental Status Examination (3MSE) scores among American Indian adults, ages 65–95 years

Model	Age γ (SE)	Cohen's d	Sex γ (SE)	Cohen's d	Education γ (SE)	Cohen's d	Language γ (SE)	Cohen's d
VEM	−0.09 (0.05)	0.18	−0.06 (0.03)	0.12	−0.02 (0.06)	0.04	0.04 (0.04)	0.08
Overall	−0.30 (0.04)*	0.63	−0.02 (0.04)	0.04	0.28 (0.06)*	0.58	−0.14 (0.04)*	0.28
LEF	−0.01 (0.05)	0.02	0.10 (0.03)*	0.20	0.15 (0.06)*	0.30	−0.15 (0.04)*	0.30
Overall	−0.32 (0.05)*	0.68	−0.10 (0.04)*	0.20	0.20 (0.04)*	0.41	−0.08 (0.04)	0.16
PMWM	−0.01 (0.08)	0.02	−0.17 (0.06)*	0.35	0.04 (0.07)	0.08	0.09 (0.06)	0.18
Overall	−0.32 (0.04)*	0.68	−0.02 (0.04)	0.04	0.27 (0.05)*	0.56	−0.14 (0.04)*	0.28
OVC	0.10 (0.06)	0.20	0.09 (0.03)*	0.18	−0.17 (0.05)*	0.35	0.10 (0.04)*	0.20
Overall	−0.33 (0.04)*	0.70	−0.07 (0.04)	0.14	0.29 (0.04)*	0.61	−0.16 (0.03)*	0.32

Note: * $p < 0.05$ after Benjamini-Hochberg adjustment. SE: Standard Error; VEM: Verbal Episodic Memory; LEF: Language & Executive Functioning; PMWM: Psychomotor & Working Memory; OVC: Orientation & Visuo-construction.

Table 6. Multiple Indicator Multiple Cause (MIMIC) step 2 regression coefficients (β) for a second-order, 4-factor structure of Modified Mini Mental Status Examination (3MSE) scores with bilingual speaking ability, adjusted for age, sex, and education, among American Indian adults, ages 65–95 years

Item	β Estimate (SE)
VEM	
Recall 1	−0.01 (0.03)
Recall 2	0.01 (0.03)
LEF	
Animal	−0.01 (0.04)
Command	0.09 (0.04)
Name	0.02 (0.04)
Obey	0.09 (0.05)
Repetition	−0.06 (0.04)
Similarities	−0.10 (0.04)
PMWM	
Reversal	0.40 (0.21)
Write	−0.12 (0.03)*
OVC	
Orientation	0.07 (0.03)
Pentagons	0.07 (0.04)
Registration	0.01 (0.03)
Spatial	−0.01 (0.03)
Temporal	−0.09 (0.04)

Note: * $p < .05$ via Benjamini-Hochberg adjustment. VEM = Verbal Episodic Memory, LEF = Language and Executive Function, PMWM = Psychomotor & Working Memory, OVC = Orientation & Visuo-construction.

Inference for extrapolation: MIMIC models

In MIMIC models (Table 5), step 1, differences in the second-order (Overall, General, or total score) and each of the 4 first-order factors (VEM, LEF, PMWM, OVC) were evaluated across sex, age, education, and bilingual strata. Collinearity among variables is unlikely to cause undue influence on our estimates, because the range of correlations between features was small: Spearman rho from 0.22 (age and bilingual status) to −0.17 (age and education). After adjustment for multiple statistical tests, significant differences were detected between sexes (male, female) for language and executive functioning (LEF), psychomotor and working memory (PMWM), and orientation visuo-construction (OVC) factors: males scored higher than females on OVC and LEF factors, but lower on PMWM, with low to moderate effect sizes. For education (≤12, 13+ years), differences were detected for LEF and OVC; those with at least 13 years of education scored higher in LEF but lower in OVC than those with up to 12 years of education, with moderate effect sizes. Bilingual ability (not at all/a little, moderately/very well) was also associated with LEF and OVC factors, wherein those who were better at speaking their Native language – and thus represented bi- or multi-lingual speakers – scored lower on LEF but higher on OVC than those who were not bilingual, with moderate and low effects, respectively. Significant associations with the total Overall factor, or total score, were consistently detected for age, wherein older participants scored lower, with high effect estimates ($d > 0.6$). Sex, education, and language were also associated with the Overall factor, with moderate effect sizes ($d = 0.2$–0.4).

In step 2 of MIMIC modeling (Table 6), differences across bilingual ability groups, independent of age, sex, and education, were only significant for the *Writing* task ($p < 0.05$). Specifically, participants with higher bilingual speaking ability had lower scores on this task; however, the effect was small (Cohen's $d = 0.24$). Bilingual speakers had similar effect sizes with lower scores on the similarities task and a larger effect size but higher scores than non-bilingual individuals on the mental reversal task; however, these differences were not statistically significant.

Discussion

This study presents data supporting some validity inferences for 3MSE scores among American Indian adults aged 65–95. Specifically, these results support the following inferences: (a) 3MSE total score can be used to represent general cognitive functioning as a summary score, a finding consistent with other populations; (b) 3MSE total score has adequate internal consistency-reliability and stability; (c) four sub-factor scores may be used to represent verbal-episodic memory, language and executive function, psychomotor skills and working memory, and orientation and visuo-construction skills, but with caution given low reliability; (d) differences in the total score were observed across strata of sex, age, and language use; in particular (e) the *Write* task differed by language use, independent of sex, age, and education. Of note, these findings suggest high degree of validity for 3MSE, despite prior reports of generally low distribution for 3MSE scores in this population (A. M. Suchy-Dicey et al., 2022; S. erney et al., 2015).

Also of note, our second-order factor structure yielded a better model fit than the unidimensional model. The more complex second-order model offers improved flexibility in test performance, given the multiple domains involved (Klyce et al., 2021), albeit not consistent with the original test design (Teng & Chang Chui, 1987). Furthermore, our high omega estimates support the inference of generalizability and use of the single summary score for individual-level interpretation of results – however, lower estimates for each of the four sub-factors caution against such interpretations for the multi-level model. Therefore, future research may be needed to justify the use of a 4-factor or other

higher-level model in individual-level interpretation in this or other similar populations – such as American Indians from other regions, from urban settings, or Alaska Native peoples.

The moderate-to-strong stability coefficients over the mean 7-year interval suggest consistent score performance over time, including across age, sex, education, and bilingual strata. However, a closer examination of stratum-specific stability estimates suggests a few groups that may warrant closer monitoring, greater score volatility, or higher potential for cognitive variance and decline, is validated for these population strata. For example, those who self-reported not using their Native language at all (i.e., English-only speakers) had the lowest score stability compared with those who spoke both English and their Native language. This finding is consistent with findings from other populations that multi-lingual speakers may have better cognitive function (Bialystok et al., 2008; Costa et al., 2008; Prior, 2010) or be protected against cognitive decline (Bialystok et al., 2007; Costumero et al., 2020; Craik et al., 2010; Crane et al., 2010; Guzman-Velez & Tranel, 2015; Ramakrishnan et al., 2017) perhaps via enhanced cognitive reserve,(Liu & Wu, 2021) especially from diseases that preferentially affect the temporal lobe (Schweizer et al., 2012)

However, those with more education also had lower score stability than those with less education, which is inconsistent with expectation; however, education in this cohort was strongly confounded by traumatic childhood experiences at Boarding schools, and has been shown to have poor construct validity for baseline cognitive function in this population (A. M. Suchy-Dicey et al., 2022), altogether emphasizing the need for more research on crystallized cognition and education as contextual factors in cognitive assessments.

MIMIC models found education to have the largest number of significant group differences, independent of other key features (age, sex, and language use); although each of the other features also had significant, mutually-independent findings for both the four factor- and general factor models. These findings suggest that all of these features – age, sex, education, and language use – may be important for interpreting both overall and sub-factor scores on 3MSE. Differential item functioning analysis, looking more deeply at differences across language use categories, identified one item (*Write,* wherein participants are asked to write a short phrase for up to five points) with a significant, albeit small effect. The impact of language use on test and language-specific skills suggests that score interpretation in culturally and educationally diverse settings may need to carefully consider the scoring impact of items with a language or writing component. More than one-third of our study population self-identifies as at least moderately bilingual, suggesting that test scores could be artificially reduced by up to 5% for a substantial portion of this population. Future analyses should examine whether adaptations or adjustments of the test structure or scoring matrix can improve score accuracy compared to other populations.

Because the 3MSE was not designed with American Indian or other culturally-diverse populations in mind, test validation is vital to score interpretation in these groups. Given that educational, linguistic, socioeconomic, cultural, and historical factors can influence test administration and performance – independent of cognition – extra care must be taken when administering tests to minoritized populations to limit further influences of testing artifacts. As an example, not all participants are able or willing to undergo testing in the office environment; if possible, tests can be given at home or in other secure settings, as long as the test environment is secure, quiet, and free of distractions. Because many participants may have limited access to regular healthcare, vision, and hearing limitations may be inadequately addressed, so a range of reading glasses and basic hearing assistive devices may be needed. Trust and rapport are important in community settings, and many interviewers are younger than the individual Elders being assessed, which can pose a cultural difficulty in correcting errors during testing. Finally, not all American Indian populations and groups are homogeneous or similar in history, language, culture, and other contextual factors, so knowledge of the environment and population of interest, with staff, clinic, and physician training in local perspectives, experiences, and histories may be important to achieving quality elder care.

This work must be evaluated in context with methodological considerations. First, these data comprise a survival cohort, with data from over 800 individuals at Visit 1 but only around 400 at Visit 1 and 2. Thus, inferences related to reliability, internal consistency and stability, were limited to those who could attend both visits and may be biased toward healthier inclusion. In addition, although our results support scoring and interpretation of 3MSE scores, the implications of whether these scores represent normal or impaired cognition and what scores would provide accurate and precise categorization are not yet addressed with these data; future research with physician panel based case review data will address these questions about the utility of 3MSE score use in practice (Cizek et al., 2008). Also, our data include a large, heterogeneous cohort of American Indian participants from 11 tribes and communities over 3 geographic regions; however, other groups may not be represented by these groups and may require independent score validity assessments.

Overall, formal psychometric analysis to interpret 3MSE scores, overall and in the context of factors such as age, sex, education, and language use, is needed for valid, appropriate inference. A growing body of evidence suggests that standards from White populations do not generalize to others, suggesting a potential need for independent validations in all peoples. These data confirm the generally valid use of 3MSE scores in American Indian individuals but with a possible need for score contextualization by sex, education, and language use (bilingual) strata.

References

Avila, J. F., Renteria, M. A., Witkiewitz, K., Verney, S. P., Vonk, J. M. J., & Manly, J. J. (2020). Measurement invariance of neuropsychological measures of cognitive aging across race/ethnicity by sex/gender groups. *Neuropsychology*, 34(1), 3–14. https://doi.org/10.1037/neu0000584

Bassuk, S. S., & Murphy, J. M. (2003). Characteristics of the modified mini-mental state exam among elderly persons. *Journal of Clinical Epidemiology*, 56(7), 622–628. https://doi.org/10.1016/S0895-4356(03)00111-2

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, *45*(2), 459–464. https://doi.org/10.1016/j.neuropsychologia.2006.10.009

Bialystok, E., Craik, F. I. M., Luk, G. (2008). Lexical access in bilinguals: Effects of vocabulary size and executive control. *Journal of Neurolinguistics*, *21*(6), 522–538.

Brave Heart, M. Y., & DeBruyn, L. M. (1998). The American Indian Holocaust: Healing historical unresolved grief. *American Indian and Alaska Native Mental Health Research*, *8*(2), 56–78.

Bravo, G., & Hébert, R. (1997). Age- and education-specific reference values for the Mini-Mental and modified Mini-Mental State Examinations derived from a non-demented elderly population. *International Journal of Geriatric Psychiatry*, *12*(10), 1008–1018. https://doi.org/10.1002/(sici)1099-1166(199710)12:

Brickman, A. M., Cabo, R., & Manly, J. J. (2006). Ethical issues in cross-cultural neuropsychology. *Applied Neuropsychology*, *13*(2), 91–100. https://doi.org/10.1207/s15324826an1302_4

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd edn). The Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrd, D. A., & Rivera-Mindt, M. G. (2022). Neuropsychology's race problem does not begin or end with demographically adjusted norms. *Nature Reviews Neurology*, *18*(3), 125–126. https://doi.org/10.1038/s41582-021-00607-4

Cappeliez, P., Quintal, M., Blouin, M., Gagne, S., Bourgeois, A., Finlay, M., & Robillard, E. T. A. (1996). Les Propriétés Psychométriques De La Version Française Du Modified Mini-Mental State (3Ms) Avec Des Patients âgés Suivis En Psychiatrie Gŕriatrique. *The Canadian Journal of Psychiatry*, *41*(2), 114–121. https://doi.org/10.1177/070674379604100209

Census (2020). *United States Census*. United States Census Bureau. Retrieved from census.gov.

Cizek, G., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397, 412. https://doi.org/10.1177/0013164407310130

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd edn). L. Erlbaum Associates.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Costa, A., Hernandez, M., & Sebastian-Galles, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*(1), 59–86. https://doi.org/10.1016/j.cognition.2006.12.013

Costumero, V., Marin-Marin, L., Calabria, M., Belloch, V., Escudero, J., Baquero, M., Hernandez, M., de Miras, J.R., Costa, A., Parcet, M. A., & Avila, C. (2020). A cross-sectional and longitudinal study on the protective effect of bilingualism against dementia using brain atrophy and cognitive measures. *Alzheimers Research & Therapy*, *12*(11). https://doi.org/10.1186/s13195-020-0581-1.

Craik, F. I., Bialystok, E., & Freedman, M. (2010). Delaying the onset of Alzheimer disease: Bilingualism as a form of cognitive reserve. *Neurology*, *75*(19), 1726–1729. https://doi.org/10.1212/WNL.0b013e3181fc2a1c

Crane, P. K., Gruhl, J. C., Erosheva, E. A., Gibbons, L. E., McCurry, S. M., Rhoads, K., Nguyen, V., Arani, K., Masaki, K., White, L. (2010). Use of spoken and written Japanese did not protect Japanese-American men from cognitive decline in late life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *65*(6), 654–666. https://doi.org/10.1093/geronb/gbq046

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA*, *269*(18), 2386–2391, https://doi.org/10.1001/jama.1993.03500180078038.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*(3), e0121945. https://doi.org/10.1371/journal.pone.0121945

Fernandes, M., Ayede, A. I., & Blackmon, K. (2022). Addressing racial inequities in neuropsychological assessment requires international prescriptive standards, not demographically adjusted norms. *Nature Reviews Neurology*, *18*(6), 377–377. https://doi.org/10.1038/s41582-022-00652-7

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In *Structural equation modeling: A second course* (2nd ed. pp. 439–492). IAP Information Age Publishing.

Folstein, M. F., Robins, L. N., & Helzer, J. E. (1983). The mini-mental state examination. *Archives of General Psychiatry*, *40*(7), 812. https://doi.org/10.1001/archpsyc.1983.01790060110016

Galvin, J. E., Chrisphonte, S., & Chang, L. C. (2021). Medical and social determinants of brain health and dementia in a multicultural community cohort of older adults. *Journal of Alzheimers Disease*, *84*(4), 1563–1576. https://doi.org/10.3233/JAD-215020

Grace, J., Nadler, J. D., White, D. A., Guilmette, T. J., Giuliano, A. J., Monsch, A. U., & Snow, M. G. (1995). Folstein vs modified Mini-Mental State Examination in geriatric stroke. Stability, validity, and screening utility. *Arch Neurol*, *52*(5), 477–484. https://doi.org/10.1001/archneur.1995.00540290067019

Guzman-Velez, E., & Tranel, D. (2015). Does Bilingualism contribute to cognitive reserve? Cognitive and neural perspectives. *Neuropsychology*, *29*(1), 139–150. https://doi.org/10.1037/neu0000105

Jervis, L. L., Beals, J., Fickenscher, A., & Arciniegas, D. B. (2007). Performance on the mini-mental state examination and mattis dementia rating scale among older American Indians. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *19*(2), 173–178. https://doi.org/10.1176/jnp.2007.19.2.173

Jervis, L. L., Fickenscher, A., Beals, J., Cullum, C. M., Novins, D. K., Manson, S. M., & Arciniegas, D. B. (2010). Predictors of performance on the MMSE and the DRS-2 among American Indian elders. *Journal of Neuropsychiatry*, *22*(4), 417–425. https://doi.org/10.1176/appi.neuropsych.22.4.417

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kim, G., Bryant, A. N., Goins, R. T., Worley, C. B., & Chiriboga, D. A. (2012). Disparities in health status and health care access and use among older American Indians and Alaska Natives and non-Hispanic Whites in California. *Journal of Aging and Health*, *24*(5), 799–811. https://doi.org/10.1177/0898264312444309

Klyce, D. W., Pugh, M., & Perrin, P. B. (2021). Factor structure of the Modified Mini Mental State (3MS) exam among participants in inpatient neuro-rehabilitation. *Brain Injury*, *35*(11), 1425–1432. https://doi.org/10.1080/02699052.2021.1972449

Lee, E. T., Welty, T. K., Fabsitz, R., Cowan, L. D., Le, N. A., Oopik, A. J., Cucchiara, A., Savage, P., & Howard, B. V. (1990). The Strong Heart Study. A study of cardiovascular disease in American Indians: Design and methods. *American Journal of Epidemiology*, *132*(6), 1141–1155. https://doi.org/10.1093/oxfordjournals.aje.a115757

Lenhard, W., & Lenhard, A. (2016). *Computation of effect sizes*,

Lezak, M. D. (2012). *Neuropsychological assessment* (5th edn). Oxford University Press.

Liu, H., & Wu, L. (2021). Lifelong bilingualism functions as an alternative intervention for cognitive reserve against Alzheimer's disease. *Frontiers in Psychiatry*, *12*, 696015. https://doi.org/10.3389/fpsyt.2021.696015

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., Orgeta, V., Ritchie, K., Rockwood, K., Sampson, E. L., Samus, Q., Schneider, L. S., Selbæk, G., Teri, L., Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*, *396*(10248), 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Manly, J. J., Jacobs, D. M., Sano, M., Bell, K., Merchant, C. A., Small, S. A., & Stern, Y. (1998). Cognitive test performance among nondemented elderly African Americans and whites. *Neurology*, *50*(5), 1238–1245. https://doi.org/10.1212/wnl.50.5.1238

McCann, A., McNulty, H., Rigby, J., Hughes, C. F., Hoey, L., Molloy, A. M., Cunningham, C. J., Casey, M. C., Tracey, F., O'Kane, M. J., McCarroll, K., Ward, M., Moore, K., Strain, J. J., Moore, A. (2018). Effect of area-level socioeconomic deprivation on risk of cognitive dysfunction in older adults.

Journal of The American Geriatrics Society, *66*(7), 1269–1275. https://doi.org/10.1111/jgs.15258

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

McDowell, I., Kristjansson, B., Hill, G. B., & Hébert, R. (1997). Community screening for dementia: The Mini Mental State Exam (MMSE) and modified Mini-Mental State Exam (3MS) compared. *Journal of Clinical Epidemiology*, *50*(4), 377–383. https://doi.org/10.1016/S0895-4356(97)00060-7

Mehta, K. M., & Yeo, G. W. (2017). Systematic review of dementia prevalence and incidence in United States race/ethnic populations. *Alzheimers & Dementia*, *13*(1), 72–83. https://doi.org/10.1016/j.jalz.2016.06.2360

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.

Nunnally, J. C. B. I. H. (1994). *Psychometric theory*. McGraw-Hill.

Petrill, S. A., Luo, D., Thompson, L. A., & Detterman, D. K. (1996). The independent prediction of general intelligence by elementary cognitive tasks: Genetic and environmental influences. *Behavior Genetics*, *26*(2), 135–147. https://doi.org/10.1007/BF02359891

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage Pub.

Prior, A. M., & B (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, *13*(2), 253–262. https://doi.org/10.1017/S1366728909990526

Ramakrishnan, S., Mekala, S., Mamidipudi, A., Yareeda, S., Mridula, R., Bak, T. H., Alladi, S., Kaul, S. (2017). Comparative effects of education and bilingualism on the onset of mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, *44*(3-4), 222–231. https://doi.org/10.1159/000479791

Rapp, S. R., Espeland, M. A., Hogan, P., Jones, B. N., & Dugan, E. (2003). Baseline experience with Modified Mini Mental State Exam: The Women's Health Initiative Memory Study (WHIMS). *Aging & Mental Health*, *7*(3), 217–223. https://doi.org/10.1080/1360786031000101201

Ryan, J., Woods, R. L., Britt, C., Murray, A. M., Shah, R. C., Reid, C. M., Kirpach, B., Wolfe, R. S., Nelson, M. R., Lockery, J. E., Orchard, S. G., Trevaks, R. E., McNeil, J. J., Storey, E. (2019). Normative performance of healthy older individuals on the Modified Mini-Mental State (3MS) examination according to ethno-racial group, gender, age, and education level. *Clinical Neuropsychologist*, *33*(4), 779–797. https://doi.org/10.1080/13854046.2018.1488996

Schweizer, T. A., Ware, J., Fischer, C. E., Craik, F. I., & Bialystok, E. (2012). Bilingualism as a contributor to cognitive reserve: Evidence from brain atrophy in Alzheimer's disease. *Cortex*, *48*(8), 991–996. https://doi.org/10.1016/j.cortex.2011.04.009

Sink, K. M., Craft, S., Smith, S. C., Maldjian, J. A., Bowden, D. W., Xu, J., Freedman, B. I., Divers, J. (2015). Montreal cognitive assessment and modified mini mental state examination in African Americans. *Journal of Aging Research*, *2015*, 872018–6. https://doi.org/10.1155/2015/872018.

Suchy-Dicey, A., Oziel, K., Sawyer, C., Olufadi, Y., Ali, T., Fretts, A. M., & Umans, J. G., Shibata, D. K., Longstreth, W. T., & Grabowski, T. J. (2022). Educational and clinical associations with longitudinal changes in cognitive function and brain imaging among American Indians: the Strong Heart Study. *Neurology*, *99*(24), e2637–e2647.

Suchy-Dicey, A., Verney, S. P., Nelson, L. A., Barbosa-Leiker, C., Howard, B. A., Crane, P. K., & Buchwald, D. S. (2020). Depression symptoms and cognitive test performance in older American Indians: The strong heart study. *Journal of The American Geriatrics Society*, *68*(8), 1739–1747. https://doi.org/10.1111/jgs.16434

Suchy-Dicey, A. M., Oziel, K., Sawyer, C., Olufadi, Y., Ali, T., Fretts, A. M., Umans, J. G., Shibata, D. K., Longstreth, W. T., Rhoads, K., Buchwald, D. S., & Grabowski, T. J. (2022). Educational and clinical associations with longitudinal cognitive function and brain imaging in American Indians: The strong heart study. *Neurology*, *99*(24), e2637–2647. https://doi.org/10.1212/WNL.0000000000201261

Suchy-Dicey, A. M., Shibata, D., Best, L. G., Verney, S. P., Longstreth, W. T., Lee, E. T., Okin, P. M., Devereux, R., O'Leary, M., Ali, T., Jensen, P. N., Muller, C., Nelson, L. A., Rhoades, E., Madhyastha, T., Grabowski, T. J., Beauchamp, N., Umans, J. G., Buchwald, D. (2016). Cranial magnetic resonance imaging in elderly American Indians: Design, methods, and implementation of the cerebrovascular disease and its consequences in American Indians study. *Neuroepidemiology*, *47*(2), 67–75. https://doi.org/10.1159/000443277

Suchy-Dicey, A. M., Shibata, D. K., Madhyastha, T. M., Grabowski, T. J., Longstreth, W. T., & Buchwald, D. S. (2017). Findings of vascular brain injury and structural loss from cranial magnetic resonance imaging in elderly American Indians: The strong heart study. *Neuroepidemiology*, *48*(1-2), 39–47. https://doi.org/10.1159/000459624

Teng, E. L., & Chang Chui, H. (1987). The modified mini-mental (3MS) examination. *Journal Clinical Psychiatry*, *48*(8), 314–318.

Tombaugh, T. N. (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Archives of Clinical Neuropsychology*, *20*(4), 485–503. https://doi.org/10.1016/j.acn.2004.11.004

Tombaugh, T. N., McDowell, I., Kristjansson, B., & Hubley, A. M. (1996). Mini-Mental State Examination (MMSE) and the Modified MMSE (3MS): A psychometric comparison and normative data. *Psychological Assessment*, *8*(1), 48–59. https://doi.org/10.1037/1040-3590.8.1.48.

Tschanz, J. T., Welsh-Bohmer, K. A., Plassman, B. L., Norton, M. C., Wyse, B. W., & Breitner, J. C. (2002). An adaptation of the modified mini-mental state examination: Analysis of demographic influences and normative data: The cache county study. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *15*(1), 28–38.

Tsoi, K. K., Chan, J. Y., Hirai, H. W., Wong, S. Y., & Kwok, T. C. (2015). Cognitive tests to detect dementia: A systematic review and meta-analysis. *JAMA Internal Medicine*, *175*(9), 1450–1458. https://doi.org/10.1001/jamainternmed.2015.2152

Van Patten, R., Britton, K., & Tremont, G. (2019). Comparing the Mini-Mental State Examination and the modified Mini-Mental State Examination in the detection of mild cognitive impairment in older adults. *International Psychogeriatrics*, *31*(5), 693–701. https://doi.org/10.1017/S1041610218001023

Verney, S., Bennett, J., & Hamilton, J. M. (2015). Cultural considerations in the neuropsychological assessment of American Indians/Alaska natives. In *Minority and cross-cultural aspects of neuropsychological sssessment: Enduring and emerging trends* (2nd ed. pp. 115–158).

Verney, S. P., Suchy-Dicey, A. M., Cholerton, B., Calhoun, D., Nelson, L., Montine, T. J., Ali, T., Longstreth, W. T., Buchwald, D. (2019). The associations among sociocultural factors and neuropsychological functioning in older American Indians: The Strong Heart Study. *Neuropsychology*, *33*(8), 1078–1088. https://doi.org/10.1037/neu0000574

Zahodne, L. B., Sharifian, N., Kraal, A. Z., Zaheed, A. B., Sol, K., Morris, E. P., Schupf, N., Manly, J. J., Brickman, A. M. (2021). Socioeconomic and psychosocial mechanisms underlying racial/ethnic disparities in cognition among older adults. *Neuropsychology*, *35*(3), 265–275. https://doi.org/10.1037/neu0000720

Zahodne, L. B., Sol, K., & Kraal, Z. (2019). Psychosocial pathways to racial/ethnic inequalities in late-life memory trajectories. *The Journals of Gerontology: Series B*, *74*(3), 409–418. https://doi.org/10.1093/geronb/gbx113