HHS Public Access

Author manuscript

Assessment. Author manuscript; available in PMC 2024 April 01.

Published in final edited form as:

Assessment. 2024 April ; 31(3): 745–757. doi:10.1177/10731911231180127.

Psychometric Properties of Controlled Oral Word Association (COWA) Test and Associations With Education and Bilingualism in American Indian Adults: The Strong Heart Study

Astrid M. Suchy-Dicey[1], **Thao T. Vo**[1], **Kyra Oziel**[1], **Roxanna King**[1], **Celestina Barbosa-Leiker**[1], **Kristoffer Rhoads**[2], **Steven Verney**[3], **Dedra S Buchwald**[1], **Brian F. French**[1]

[1]Washington State University, Seattle, USA

[2]University of Washington, Seattle, USA

[3]The University of New Mexico, Albuquerque, USA

Abstract

The Controlled Oral Word Association (COWA) test is used to assess phonemic fluency and executive function. Formal validation of test scores is important for accurate cognitive evaluation. However, there is a dearth of psychometric validation among American Indian adults. Given high burden of dementia risk and key contextual factors associated with cognitive assessments, this represents a critical oversight. In a large, longitudinal population-based cohort study of adult American Indians, we examined several validity inferences for COWA, including scoring, generalization, and extrapolation inferences, by investigation of factor structure, internal consistency, test–retest reliability, and differential test functioning. We found adequate unidimensional model fit, with high factor loadings. Internal consistency reliability and test–retest reliability were 0.88 and 0.77, respectively, for the full group. COWA scores were lowest among the oldest, lowest education, bilingual speakers; group effects for sex and bilingual status were small; age effect was medium; and education effect was largest. However, Wide Range Achievement Test (WRAT) score effect was stronger than education effect, suggesting better contextualization may be needed. These results support interpretation of total COWA score, including across sex, age, or language use strata.

Keywords

verbal fluency; validity; reliability; differential item functioning; American Indian

Standardized cognitive assessments rely on formal psychometric analysis for accurate, precise test interpretability and to support a validity argument in clinical applications (Lezak, 2012). Psychometric assessments, which may include validity, reliability, and valuation of test standards for diagnostics, are well-established among majority populations such as U.S.

Corresponding Author: Astrid M. Suchy-Dicey, Elson S Floyd College of Medicine, Washington State University, 1100 Olive Way Suite 1200, Seattle, WA 98101, USA. astrid.suchy-dicey@wsu.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

non-Hispanic Whites (NHW) for most tests, including the Controlled Oral Word Association (COWA) test (Tombaugh et al., 1999). However, many minoritized populations, such as American Indians, have not been systematically included in psychometric evaluations of most standard tests, and when minority groups differ meaningfully from the majority in either diagnostic or normative standards, such oversights are likely to preclude clinicians and researchers from making valid diagnostic inference (Kiselica et al., 2021; Manly et al., 1998; Pedraza et al., 2012; Rivera Mindt et al., 2010). Given high risk of vascular and Alzheimer's-related cognitive injury in American Indians (Mehta & Yeo, 2017; Suchy-Dicey et al., 2021; Suchy-Dicey, Howard, et al., 2022; Zhang et al., 2008), and clear disparities in cognitive test performance, compared with the majority population (Suchy-Dicey, Verney, et al., 2020; Suchy-Dicey et al., 2022; Verney et al., 2019), formal psychometric evaluation of cognitive tests is critically needed for this vulnerable population.

COWA normative test data are available for NHW aged 16 to 90 years, with standards accounting for age, sex, and education strata (Harvey & Siegert, 1999; Rodriguez-Aranda & Martinussen, 2006; Ruff et al., 1996). Some normative studies of COWA have identified no associations with age (Axelrod & Henry, 1992; Bolla et al., 1990; Ruff et al., 1996), although modest negative correlations with older age have been documented (Libon et al., 1994). Sex differences have been reported, with females performing better than males (Bolla et al., 1990; Ruff et al., 1996), although the evidence is mixed (Boone, 1999; Saykin et al., 1995; Tombaugh et al., 1999). Longer tenure of formal education has also been associated with better COWA performance (Ruff et al., 1996; Selnes et al., 1991; Tombaugh et al., 1999; Verney et al., 2019). In one sample of cognitively normal, English-speaking U.S. residents, higher education, but neither age nor sex, were associated with test performance in preliminary analyses (Ruff et al., 1996); however, a three-way analysis of variance identified that sex as an effect modifier for the relationship between education and COWA test performance, with women scoring better in the highest educational group and men scoring poorer in the lowest educational group. Despite the large bolus of data available, no single study has produced COWA normative data across all age, sex, and education strata; instead, meta-analyses combined findings from multiple study samples to produce metanorms (Adesope et al., 2010; Barry et al., 2008; Loonstra et al., 2001).

COWA Performance in American Indians

Sociodemographic features have been identified as independently associated with COWA test performance in American Indians (Verney et al., 2019). The Cerebrovascular Disease and its Consequences in American Indians Study (CDCAI), an ancillary cohort within the 30-year Strong Heart Study (SHS) representing community-dwelling tribal members across four states, identified younger age, female sex, longer educational tenure, higher income, and decreased bilingual capacity as positively associated with higher scores on COWA among American Indian elders aged 65–95. A steep decline in scores was observed after 75 years of age. Summary cognitive test scores for this cohort suggest that cognitive test score distributions are overall lower in American Indian adults compared with majority populations (Suchy-Dicey, Verney, et al., 2020), with study medians near or below conventional thresholds used to define cognitive impairment or dementia in NHW (Adesope et al., 2010; Barry et al., 2008; Loonstra et al., 2001). Whether a

majority of individuals score proportionally lower than their peers from the majority population due to the influence of sociodemographic and health disparities, or whether some unidentified subset of individuals with clinically significant cognitive syndromes is lowering the population-wide estimates is yet unknown. However, the prevalence of subjective cognitive impairment is present in only 17% of American Indian elders (Centers for Disease Control and Prevention, 2019), suggesting that conventional diagnostic thresholds have poor specificity for subjective symptoms. This disparity highlights a clear and urgent need for formal psychometric evaluation of standardized cognitive tests, with the concomitant establishment of population-relevant, appropriate score thresholds for valid and precise test score interpretations for encounters with members of this population.

Sociocultural Disparities in COWA Performance

Multiple factors may be needed to contextualize cognitive test performance, most commonly age, sex, and education (Lezak, 2012; Tombaugh et al., 1999). Of these, education may have poorer contextual relevance in American Indian elders' test performance than for other populations, such as NHW. Education is a proxy variable for baseline function, or crystallized function, which is not expected to change substantively after childhood; however, in American Indians, tenure of formal education is poorly correlated with crystallized intelligence (Suchy-Dicey et al., 2022), suggesting that formal educational experiences were of varying quality and/or that many American Indian elders obtained substantive learning outside of the classroom (Sayegh et al., 2014). The first U.S. Federal standards against forced removal of children for placement into residential schools or foster care were not established until the Indian Child Welfare Act in 1976 (Lomawaima & McCarty, 2006; Lynch, 1990), with a legacy of Indigenous trauma (Running Bear et al., 2019), depression, maladaptive behaviors (Enoch & Albaugh, 2017; Lomawaima & McCarty, 2006), cultural and linguistic losses (Lynch, 1990), and premature illness and mortality (Jack & Secwepemc Cultural Education Society, 2000) affecting a majority of American Indians now as young as middle age. In addition to higher likelihood of traumatic educational experiences, American Indians are also more likely to have shorter educational tenure, although this may be counterbalanced with extracurricular learning that is not captured in conventional measures of educational tenure (Mervis, 2009; Scribner & Cole, 1973). Therefore, although years of formal education is overall low, performance on tests related to baseline function, or crystallized intelligence, is generally high (Suchy-Dicey et al., 2022). Thus, psychometric evaluation of COWA, accounting for baseline function, comparing either direct measure of crystallized intelligence or its conventional proxy measure of educational tenure, is needed for valid score contextualization.

Language skills, such as capacity and frequency of use of other languages, are also vital to the interpretation of tests of verbal or phonemic fluency and executive function, such as COWA. Bilingual status has been associated with better executive function (Bialystok et al., 2008; Costa et al., 2008; Prior, 2010) and may protect against cognitive decline (Bialystok et al., 2007; Craik et al., 2010; Crane et al., 2010) via improved working memory and functional connectivity. In contrast, managing multiple languages, especially under pressure such as during standardized testing, may increase cognitive load, posing a disadvantage in efficient information processing (Adesope et al., 2010; Liu & Wu, 2021). Data from

the SHS and CDCAI study suggest that higher bilingual capacity is inversely associated with test performance on COWA (Verney et al., 2019), supporting the hypothesis of higher cognitive load with delay in information processing for those reporting frequent use of other languages. Furthermore, because bilingualism is closely connected to identity and expression and because childhood educational experiences may have influenced language expression and learning, frequent or proficient use of native language, in addition to English, is likely to pose a critical factor for interpretation and contextualization of verbal skill test performance in this population.

Knowledge Gaps and Objectives of Our Study

Despite the high prevalence of health disparities influencing cognitive status and the importance of contextual factors in cognitive test performance, most cognitive tests, including the COWA, have not been formally psy- chometrically evaluated in Indigenous U.S. populations, including American Indians (Verney et al., 2019). Group differences exist by age, sex, and education, with some factor interactions, in NHW; these factors—as well as bilingual status—are likely to serve as key contextual features for American Indians. It is unknown to what degree population disparities in overall population scores, or association patterns with contextual features, might persist if contributing sociodemographic and health disparities and historical traumas were removed; however, until public health programs can achieve health equity, population-specific score interpretation may be needed.

Our study aims to formally establish psychometric bases for using and interpreting COWA scores among American Indians aged 65–95 years, with key contextual features including age, sex, baseline cognitive function, and bilingual status (Kane, 2013). For validity, we examined several inferences within a use and interpretation validity framework (Lissitz, 2009). We examine a scoring inference to evaluate whether COWA test scores accurately represent a differential endorsement of the cognitive domain measured. For this, we conduct factor analysis to examine scores across different levels of functioning, with an underlying model consistent with the theory of assessing verbal fluency. For generalizability, we examine whether the conditions of observation represent universal sampling conditions for the cognitive domain assessed; for this, we evaluate internal consistency and test–retest reliability. For extrapolation, we assess whether test score performance accurately represents individual functioning; for this, we evaluate contextual features including age, sex, education and baseline cognitive function, and bilingual status. Differential functioning of subscores across strata is used to assess equality of measurement properties, controlling for latent ability. In this work, we focus on groups with varying language proficiency to highlight group differences resulting from language proficiency versus test interpretation.

Investigating measurement properties, including inferences for score validity, generalizability, and extrapolation of the COWA will help establish sex, age, education, and language contributions to score performance and a basis for overall score interpretation in American Indian elders. The findings from this study will provide the first estimates of the psychometric validity of COWA in this population, providing a better understanding of this and all cognitive test assessments to improve research and clinical evaluation in this vulnerable population.

Method

Setting and Participants

From 1989 to 1991, the SHS recruited middle-aged individuals from any of 13 partnering tribes and communities in the Northern Plains, Southern Plains, and Southwest United States (Lee et al., 1990). In 2010–2013, when SHS participants were 65–95 years old, all surviving participants were invited to undergo detailed cognitive testing, brain imaging, and clinical examinations as part of the SHS ancillary CDCAI study (Suchy-Dicey et al., 2016); these examinations were repeated in 2017–2019, with the addition of measures related to Alzheimer's disease (Suchy-Dicey et al., 2022). Every effort was made for complete cohort ascertainment, with >85% recruitment of survivors at CDCAI Visit 1 (N = 817) and >75% at CDCAI Visit 2 (N = 403). Evaluation of survival bias from original to CDCAI Visit 1 based on cardiovascular risk features and events suggested that participants were younger and healthier than those who did not survive but not significantly different in vascular risk (Suchy-Dicey et al., 2018). Other reports from this cohort have characterized socioeconomic, language, and cultural features in relation to cognition in this population (Suchy-Dicey et al., 2016, 2022; Suchy-Dicey, Eyituoyo, et al., 2022; Suchy-Dicey, Verney, et al., 2020; Verney et al., 2019). All participating tribal review boards, Indian Health Service, and institutional review boards approved study activities; all participants provided written informed consent.

Data Collection Procedures

Both SHS-CDCAI examination visits involved detailed questionnaires and cognitive testing with identical protocols. Field centers conducted the examinations, with the cognitive tests and interviews administered by personnel specially trained for these neuropsychological examinations. Quality control efforts included regular feedback by clinical neuropsychologists in test administration procedures. Cognitive test instruments were first double-scored, then brought to an adjudication panel consisting of the project lead, epidemiologist, neuropsychologist, and all research coordinator personnel for open discussion and concurrence. The same individuals scored all instruments at both examinations. All data were double-entered into study databases for data verification.

COWA.—The COWA test, one of two commonly used versions which use either letters F, A, S or C, F, L is designed to provide a score reflective of verbal or phonemic fluency and executive function (Barry et al., 2008). Participants are asked to generate and pronounce as many English words as possible beginning with each of the letters F, A, and S in three serial trials or tasks lasting 1 minute each. Proper nouns, numbers, and repeated words are not allowed, and considered errors. The number of words produced for each letter is summarized by total and correct responses within and across the three letters. Previous studies of COWA document internal consistency reliability with alpha coefficients ranging from .87 to .91 (Bassuk & Murphy, 2003; McDowell et al., 1997; Tombaugh et al., 1996), and Pearson's test–retest stability coefficients ranging from .78 to .85 (Bassuk & Murphy, 2003; Grace et al., 1995).

Other Measures.—Participant's self-reported age (60–69, 70–79, 80 + years), sex, years of formal education (12 years or fewer, 13 or more years), and ability to speak their native language (not at all, a little, moderately well, very well). Native language speaking capacity is also an index of bilingual status as all participants are fluent in English as a requirement for participation in SHS. Wide Range Achievement Test Version 4 reading test (WRAT), a measure of achievement and crystallized intelligence, is a sensitive proxy of premorbid function and is not expected to decline or change substantively throughout adulthood (Casaletto et al., 2015; Veizel & Zibulsky, 2013).

Data Analyses

Validity of the Unidimensional Model.—COWA test structure was examined using confirmatory factor analysis (CFA) with robust maximum likelihood estimation to evaluate whether scores accurately capture the construct of verbal (phonemic) fluency and executive function. Based on standard guidance (Benton et al., 1994), COWA administrators recommend evaluating and interpreting a single, total score, suggesting a unidimensional construct. Thus, these analyses specified a one-factor model with three-letter scores. Model fit was evaluated with indices including root mean square error of approximation (RMSEA ≤ 0.08 (Brown, 2015), standardized root mean square residual (SRMR < 0.05; Muthén, 1998–2004), comparative fit index (CFI ≥ 0.90; Browne & Cudeck, 1992), and Tucker–Lewis index (TLI ≥ 0.90; Muthén, 1989). Confidence intervals are reported with the RMSEA estimate to aid fit evaluation, as the estimate can have artificially large values with models with low degrees of freedom. Even though it is suggested that the RMSEA not be used with such models (Kenny et al., 2015), the index is reported for completeness.

Generalizability: Internal Consistency Reliability.—COWA score reliability was examined based on parameter estimates from the CFA model, using the omega coefficient to examine internal consistency reliability estimates for the total sample and across age, sex, education, and language use categories (McDonald, 1999). Omega values ≥ 0.90 are considered adequate for individual decisionmaking and ≥ 0.8 adequate for research (Nunnally, 1994).

Generalizability: Test–Retest Reliability.—Pearson's correlation coefficients were used to examine test–retest reliability, comparing the two longitudinal examination visits (2010–2013, 2017–2019). Benjamini–Hochberg's false discovery rate was used to control for multiple comparisons (Benjamini & Hochberg, 1995).

Extrapolation: MIMIC Model Primary Analysis.—Multiple indicator multiple cause (MIMIC) structural equation model was used to first examine an extrapolation inference (Figure 1). Also referred as known group differences, MIMIC estimates whether total COWA scores differentiate across various grouping variables, including sex, age, education, and language ability in an expected direction. To aid interpretation of the MIMIC model, language skills were dichotomized: speaks native language (bilingual with English) either not at all or a little, versus moderately or very well. Second, differential indicator analysis (DIF) was examined across the two language ability groups for each of the three-letter tasks, controlling for sex and age. Third, for any indicator that resulted in statistical significance

from Step 2, a subsequent model was examined, with education as a mediating variable (Cheng et al., 2016). Effect sizes were calculated and converted to *d* to be on a common metric (Lenhard & Lenhard, 2016), with 0.2, 0.5, and 0.8 benchmarks representing small, medium, and large effect sizes, respectively (Lakens, 2013).

Secondary Analysis Comparing WRAT Versus Education.—Education tenure (length) is a proxy measure for baseline function, often conceptualized as crystallized intelligence. However, education tenure may not be accurate proxy for achievement, baseline function, or crystallized intelligence, with poor factor correlation (Suchy-Dicey et al., 2022). Education was used in primary analyses because it was provided by all participants and is the conventional contextual measure for cognitive evaluation. However, WRAT reading scores were also considered in secondary models. WRAT was only collected at Visit 2 and is thus available in only 403 participants. Initial data quality checks suggested WRAT scores were missing not at random among those at Visit 2 (Baraldi & Enders, 2010; Widaman, 2006), whereas age and sex are both significant predictors of missingness, as would be expected. MIMIC models comparing education with WRAT were estimated on a restricted sample of only those who attended Visit 2 (Little, 1988).

Results

Study Sample

As in previous reports (Suchy-Dicey et al., 2016; Suchy-Dicey, Shibata, et al., 2020), the study population ($N = 818$) was older and majority female (Table 1). Mean tenure education was around 12–13 years, and majority were at least moderately bilingual. Mean correct words for the COWA test (total score) was approximately 24 at each of the two visits, and mean correctly pronounced WRAT reading words (WRAT score) was approximately 41 at Visit 2.

Structure and Scoring: Unidimensional Model

The unidimensional or single-factor model of verbal fluency and executive function (e.g., COWA total correct score) using the F, A, and S tasks yielded adequate model fit: $\chi^2(1) = 9.74$, $p < .01$, CFI = 0.99, SRMR = 0.04, RMSEA = 0.10 (95% confidence interval [CI]: 0.05, 0.17). The RMSEA CI contains the value of 0.08, a model fit criterion for this study, justifying moving forward with this model. In addition, RMSEA is biased upward in models with low degrees of freedom, and CFI and SRMR are recommended in such instances (Shi et al., 2022). Given three indicators of the construct, the A and S pattern coefficients were constrained to be equal, to gain 1 degree of freedom to obtain model fit indices. In the just-identified model, the factor loadings were high, and similar across F, A, and S scores (loadings = 0.84, 0.88, and 0.82, respectively), justifying the constraint to fit indices.

Generalizability: Internal Consistency Reliability

The internal consistency reliability estimate for the overall score with the total sample was $\omega = .88$ (Table 2). Internal consistency reliability estimates were adequate across subgroups, ranging from 0.84 to 0.90, with the highest value associated with the 70-0 to 79-year-old age group and the lowest value associated with the 65- to 69-year-old age group.

Generalizability: Test–Retest Reliability

Means were similar over the two study visits for all groups (Table 3) but were lowest among those in the highest age group (above age 80), the lowest education group (with less than 12 years of education), or those who are bilingual (speak their native language [very well]). Pearson's correlation coefficients between Visit 1 and Visit 2 ranged from .70 to .80 across all pairwise comparisons of sex, age, education, and bilingual strata. The lowest reliability correlation (r = .70) was among those who speak their native language *moderately well*; the largest reliability correlation (r = .80) was among those who speak their native language *very well*. Within-group pairwise comparisons were not statistically significant.

Extrapolation: MIMIC Model Primary Analysis

In MIMIC models, results for Step 1 identified significant COWA score differences across sex, age, education, and language groups (Table 4). Sex had a small effect, with females scoring higher than males (d = 0.20). Bilingual status also had a small effect, with poorer or less bilingual ability corresponding to better COWA scores (d = 0.15). Age had a medium effect, with younger scoring higher than older (d = 0.49). Education had the largest effect, with more education corresponding to higher COWA scores (d = 0.80). In Step 2, where each letter subtest was regressed onto language ability groups to examine differential difficulty on these tasks, there were no statistically significant differences among F, A, or S scores (p > .05). Given nonsignificant results in Step 2, mediation models for Step 3 were not completed.

Secondary Analysis: WRAT Versus Education

Using a restricted sample (i.e., Visit 2 only) to compare education effect against WRAT score effect (Table 5), Step 1 model fit with education showed similar results as the full sample, demonstrating proof of concept for restricted sample. In this model, education (d = 0.096, p < .05) and age (d = 0.32, p < .05) had significant relationships with COWA total score; sex and bilingual fluency did not. Comparatively, the model with WRAT scores in place of education, variables for sex, age, and language were not statistically significant, but the effect size associated with age remained consistent (d = 0.32 vs. 0.26). In that model, WRAT scores were statistically significant, with a large effect size (d = 1.24, p < .05). In Step 2, where each letter subtest score was regressed onto language ability groups to examine differential difficulty on these tasks, controlling for overall fluency, no statistically significant differences were found for individual word (F, A, or S) scores. Given nonsignificant results in Step 2, the mediation models for Step 3 were not completed. The standardized parameter estimates using the full sample are presented in Table 5. Step 2 was only conducted once across the two models (education and WRAT) because these variables were not involved in assessing differential test effects.

Discussion

This study is investigated inferences to support the validity, generalizability, and stratum-specific use of COWA scores among American Indian adults aged 65–95 years. As most studies with the COWA have focused on NHW, this investigation provides needed evidence on the use of COWA in a racialized and historically marginalized population. Our results

support inferences that COWA can be used as a unidimensional score, that COWA total correct score has adequate internal consistency and stability reliability estimates, in accord with prior work in other populations, and that controlling for sex, age, and education, individuals with greater bilingual capacity perform less well on COWA—similar to previous findings (Verney et al., 2019). In contrast to expectation and adjusting for overall fluency, no evidence was found that scores on individual word list tasks (F, A, or S subscales) were different based on native language proficiency (i.e., bilingual status), nor evidence of differences by age or sex. However, there were differences between effect of WRAT score, a direct measure of premorbid function, compared with education, a proxy for premorbid function. In aggregate, the evidence supports the use of COWA in this population, with the caveat that some additional inferential or contextual work is needed.

Future Directions: Heterogeneity and Intersectionality

Even though some of the validity inferences were supported, key aspects of the assessment should not be overlooked. First, American Indian participants are not a singular or homogeneous group, and there may be some contexts for which these inferences may not apply. Given that key factors influencing cognitive function and testing (Zahodne et al., 2017, 2021) include age, sex, education, language, mood, and socioeconomics, those who differ meaningfully in these areas may require additional evaluation, with test norms minimally developed according to age, sex, and educational strata.

Also, research utilizing intersectional frameworks could be explored, to highlight the influence of structural and historical inequities on cognition and cognitive testing (Cole, 2009; Crenshaw, 1989; Rosenthal, 2016). Such a multilevel framework might simultaneously examine both personal (e.g., depression) and structural (e.g., discrimination) factors (Settles et al., 2016), providing researchers and clinicians with a more nuanced understanding of interlocking systems of privilege and oppression.

Educational Context

The findings comparing direct (WRAT score) and indirect (tenure of education) measures of premorbid function are important to valid score contextualization. Prior reports have detected poor correlation between these measures in this population (Suchy-Dicey et al., 2022); the current analyses build on those findings, by reporting that WRAT score outperforms education in COWA score validity and generalizability inferences. American Indian educational paths and experiences vary considerably; however, some contributing factors include extracurricular sources of learning, traumatic experiences contributing to test anxiety, and poor educational quality at many schools, altogether reducing the utility of education as a metric. Collectively, WRAT or other measure of crystallized intelligence may be more valid, precise, and accurate in quantifying premorbid function, although future research will need to psycho- metrically validate, compare, and evaluate meaningful population strata for any such measure, in this or any other population.

Community and Cultural Context

An additional, key element for interpreting cognitive test scores is the test administration context. Conducting standardized neuropsychological research in a community setting

can be challenging due to issues of procedural suitability in environments with limited prior exposure to standardized testing, limited access to specialty health care such as neuropsychology clinics, limited support for common sensory or motor limitations, and inadequate accommodation of cultural perspectives in available testing formats. Our field center teams have developed methodologies to maximize success in guiding community-based elders through such test materials. For COWA, because many elders regularly use non-English languages or may have limited structural linguistic knowledge, our field center staff worked to ensure that the rules regarding non-allowable words were clear, providing examples using a nontested letter prior to starting. Participants were allowed to set the testing pace and given a brief break between letters if needed. Consideration of physical comfort involved selection of test environment, transportation support, provision of snacks, use of technological resources such as large font screens, and careful attention to participant rapport. Finally, our teams conduct regular consultations with community members and participants as equal partners to ensure the centralization of research subjects at the center of the research paradigm.

Limitations

This work is not without limitations. First, inferences were selected based on the secondary nature of data analysis; future work may identify additional inferences to test or exploratory analyses to conduct, with novel data collection efforts. In addition, some implications are not yet addressed; for example, whether COWA scores are able to identify early cognitive decline is unaddressed due to the as-yet unavailability of a gold standard for cognitive impairment and dementia (Cizek et al., 2008). As such data become available, such analyses will provide informative guidance for clinicians in score interpretation. Third, although these analyses were structured in a manner consistent with clinical practice and with guidance from test developers, and limited to confirmatory factor analysis due to the nature of the test structure, exploratory analyses to examine different factor model structures may be of interest, when additional data become available. Finally, this work included American Indian adults aged >65 years; future analyses may expand to include other age groups, or other Indigenous or minoritized populations.

Conclusion

In conclusion, formal validation of cognitive tests such as COWA is important for the valid evaluation of cognition and dementia. This study demonstrates COWA score validity and quantifies associations with age, sex, education, and language use in American Indians, an underserved and unique population. The impact of this work includes the three inferences for a validity argument, but with explicit acknowledgment of novel assessment needed in different contexts, assessment protocols needed to account for such contexts, and the need to empirically examine additional inferences tied to decision-making, diagnostics, and intersectionality. These findings provide a novel understanding of cognitive test performance in American Indians and are intended to help guide future clinical and research efforts in vascular, Alzheimer's, and related dementias.

Acknowledgments

The authors wish to thank all Strong Heart Study staff, participants, and communities. The opinions expressed in this paper are solely the responsibility of the author(s) and do not necessarily reflect the official views of the Indian Health Service or the National Institutes of Health (NIH).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study has been funded in whole or in part with federal funds from the National Institutes of Health, including R01HL093086, P50AG005136, and K01AG057821.

References

Adesope OO, Lavin T, Thompson T, & Ungerleider C (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. Review of Educational Research, 80(2), 207–245. 10.3102/0034654310368803

Axelrod BN, & Henry RR (1992). Age-related performance on the Wisconsin card sorting, similarities, and controlled oral word association tests. Clinical Neuropsychologist, 6(1), 16–26. 10.1080/13854049208404113

Baraldi AN, & Enders CK (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48(1), 5–37. 10.1016/j.jsp.2009.10.001 [PubMed: 20006986]

Barry D, Bates ME, & Labouvie E (2008). FAS and CFL forms of verbal fluency differ in difficulty: A meta-analytic study. Applied Neuropsychology, 15(2), 97–106. 10.1080/09084280802083863 [PubMed: 18568601]

Bassuk SS, & Murphy JM (2003). Characteristics of the modified mini-mental state exam among elderly persons. Journal of Clinical Epidemiology, 56(7), 622–628. 10.1016/S0895-4356(03)00111-2 [PubMed: 12921930]

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. 10.1111/j.2517-6161.1995.tb02031.x

Benton AL, De Hamsher SK, & Sivan AB (1994). Multilingual aphasia examination (3rd. ed.). Psychological Assessment Resources.

Bialystok E, Craik FI, & Freedman M (2007). Bilingualism as a protection against the onset of symptoms of dementia. Neuropsychologia, 45(2), 459–464. 10.1016/j.neuropsychologia.2006.10.009 [PubMed: 17125807]

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

Bolla KI, Lindgren KN, Bonaccorsy C, & Bleecker ML (1990). Predictors of verbal fluency (FAS) in the healthy elderly. Journal of Clinical Psychology, 46(5), 623–628.10.1002/1097-4679(199009)46:5<623::aid-jclp2270460513>3.0.co;2-c [PubMed: 2246370]

Boone KB (1999). Neuropsychological assessment of executive functions: Impact of age, education, gender, intellectual level, and vascular status on executive test scores. In The human frontal lobes: Functions and disorders (pp. 247–260). Guilford Press.

Brown TA (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

Browne MW, & Cudeck R (1992). Alternative ways of assessing model fit. Sociological Methods & Research, 21(2), 230–258. 10.1177/0049124192021002005

Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, & Heaton RK (2015). Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. Journal of the International Neuropsychological Society, 21(5), 378–391. 10.1017/S1355617715000351 [PubMed: 26030001]

Centers for Disease Control and Prevention. (2019). Alzheimer's disease and healthy aging. Subjective cognitive decline among American Indian/Alaska native adults. https://www.cdc.gov/aging/data/infographic/2017/american-Indian-alaska-native-cognitive-decline.html

Cheng Y, Shao C, & Lathrop QN (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. Educational and Psychological Measurement, 76(1), 43–63. 10.1177/0013164415576187 [PubMed: 29795856]

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

Cole ER (2009). Intersectionality and research in psychology. American Psychologist, 64(3), 170–180. 10.1037/a0014564 [PubMed: 19348518]

Costa A, Hernandez M, & Sebastian-Galles N (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. Cognition, 106(1), 59–86. 10.1016/j.cognition.2006.12.013 [PubMed: 17275801]

Craik FI, Bialystok E, & Freedman M (2010). Delaying the onset of Alzheimer disease: Bilingualism as a form of cognitive reserve. Neurology, 75(19), 1726–1729. 10.1212/WNL.0b013e3181fc2a1c [PubMed: 21060095]

Crane PK, Gruhl JC, Erosheva EA, Gibbons LE, McCurry SM, Rhoads K, & … White L (2010). Use of spoken and written Japanese did not protect Japanese-American men from cognitive decline in late life. Journal of Gerontology: Series B, 65(6), 654–666. 10.1093/geronb/gbq046

Crenshaw KW (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1). Article 8.

Enoch MA, & Albaugh BJ (2017). Review: Genetic and environmental risk factors for alcohol use disorders in American Indians and Alaskan Natives. American Journal of Addictions, 26(5), 461–468. 10.1111/ajad.12420

Grace J, Nadler JD, White DA, Guilmette TJ, Giuliano AJ, Monsch AU, & Snow MG (1995). Folstein vs modified Mini-Mental State Examination in geriatric stroke. Stability, validity, and screening utility. Archives of Neurology, 52(5), 477–484. 10.1001/archneur.1995.00540290067019 [PubMed: 7733842]

Harvey J, & Siegert RJ (1999). Normative data for New Zealand elders on the controlled oral word association test, graded naming test, and the recognition memory test. New Zealand Journal of Psychology, 28, 124.

Jack AS, & Secwepemc Cultural Education Society. (2000). Behind closed doors: Stories from the Kamloops Indian Residential School. Secwepemc Cultural Education Society.

Kane MT (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. 10.1111/jedm.12000

Kenny DA, Kaniskan B, & McCoach DB (2015). The performance of RMSEA in models with small degrees of freedom. Sociological Methods & Research, 44, 486–507. 10.1177/0049124114543236

Kiselica AM, Johnson E, Lewis KR, & Trout K (2021). Examining racial disparities in the diagnosis of mild cognitive impairment. Applied Neuropsychology: Adult. Advance online publication. 10.1080/23279095.2021.1976778

Lakens D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontier in Psychology, 4, Article 863. 10.3389/fpsyg.2013.00863

Lee ET, Welty TK, Fabsitz R, Cowan LD, Le NA, Oopik AJ, & … Howard BV (1990). The Strong Heart Study a study of cardiovascular disease in American Indians: Design and methods. American Journal of Epidemiology, 132(6), 1141–1155. 10.1093/oxfordjournals.aje.a115757 [PubMed: 2260546]

Lenhard W, & Lenhard A (2016). Computation of effect sizes. 10.13140/RG.2.2.17823.92329

Lezak MD (2012). Neuropsychological assessment (5th ed.). Oxford University Press.

Libon DJ, Glosser G, Malamut BL, Kaplan E, Goldberg E, Swenson R, & Prouty Sands L (1994). Age, executive functions, and visuospatial functioning in healthy older adults. Neuropsychology, 8(1), 38–43. 10.1037/0894-4105.8.1.38

Lissitz RW (2009). The concept of validity: Revisions, new directions, and applications. Information Age.

Little RJA (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202. 10.1080/01621459.1988.10478722

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

Lomawaima KT, & McCarty TL (2006). Table of contents for ''To remain an Indian'': Lessons in democracy from a century of Native American education. Teachers College Press. http://www.loc.gov/catdir/toc/fy0702/2006043864.html

Loonstra AS, Tarlow AR, & Sellers AH (2001). COWAT metanorms across age, education, and gender. Applied Neuropsychology, 8(3), 161–166. 10.1207/S15324826AN0803_5 [PubMed: 11686651]

Lynch PDCM (1990). The emergence of American Indian leadership in education. Journal of American Indian Education, 29, 1–10.

Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small SA, & Stern Y (1998). Cognitive test performance among nondemented elderly African Americans and Whites. Neurology, 50(5), 1238–1245. 10.1212/wnl.50.5.1238 [PubMed: 9595969]

McDonald RP (1999). Test theory: A unified treatment. Lawrence Erlbaum.

McDowell I, Kristjansson B, Hill GB, & 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. 10.1016/S0895-4356(97)00060-7 [PubMed: 9179095]

Mehta KM, & Yeo GW (2017). Systematic review of dementia prevalence and incidence in United States race/ethnic populations. Alzheimers Dementia, 13(1), 72–83. 10.1016/j.jalz.2016.06.2360

Mervis J. (2009). Informal education: Report calls for fresh look at what happens outside school. Science, 323(5914), 572–573. 10.1126/science.323.5914.572a

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

Muthén BO (1998-2004). Mplus technical appendices.

Nunnally JC, & Bernstein IH (1994). Psychometric theory. McGraw-Hill.

Pedraza O, Clark JH, O'Bryant SE, Smith GE, Ivnik RJ, Graff-Radford NR, & … Lucas JA (2012). Diagnostic validity of age and education corrections for the Mini-Mental State Examination in older African Americans. Journal of the American Geriatricks Society, 60(2), 328–331. 10.1111/j.1532-5415.2011.03766.x

Prior AMB (2010). A bilingual advantage in task switching. Bilingualism: Language and Cognition, 13(2), 253–262. 10.1017/S1366728909990526 [PubMed: 36479004]

Rivera Mindt M, Byrd D, Saez P, & Manly J (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: A call to action. The Clinical Neuropsychologist, 24(3), 429–453. 10.1080/13854040903058960 [PubMed: 20373222]

Rodriguez-Aranda C, & Martinussen M (2006). Age-related differences in performance of phonemic verbal fluency measured by Controlled Oral Word Association Task (COWAT): A meta-analytic study. Developemental Neuropsychology, 30(2), 697–717. 10.1207/s15326942dn3002

Rosenthal L. (2016). Incorporating intersectionality into psychology: An opportunity to promote social justice and equity. American Psychologist, 71(6), 474–485. 10.1037/a0040323 [PubMed: 27571527]

Ruff RM, Light RH, Parker SB, & Levin HS (1996). Benton controlled oral word association test: Reliability and updated norms. Archives of Clinical Neuropsychology, 11(4), 329–338. https://www.ncbi.nlm.nih.gov/pubmed/14588937 [PubMed: 14588937]

Running Bear U, Thayer ZM, Croy CD, Kaufman CE, Manson SM, & Team A-S (2019). The impact of individual and parental American Indian boarding school attendance on chronic physical health of northern plains tribes. Family and Community Health, 42(1), 1–7. 10.1097/FCH.0000000000000205 [PubMed: 30431464]

Sayegh P, Arentoft A, Thaler NS, Dean AC, & Thames AD (2014). Quality of education predicts performance on the Wide Range Achievement Test-4th Edition word reading subtest. Archives of Clinical Neuropsychology, 29(8), 731–736. 10.1093/arclin/acu059 [PubMed: 25404004]

Saykin AJ, Gur RC, Gur RE, Shtasel DL, Flannery KA, Mozley LH, & … Mozley PD (1995). Normative neuropsychological test performance: Effects of age, education, gender and ethnicity. Applied Neuropsychology, 2(2), 79–88. 10.1207/s15324826an0202_5 [PubMed: 16318528]

Scribner S, & Cole M (1973). Cognitive consequences of formal and informal education: New accommodations are needed between school-based learning and learning experiences of everyday life. Science, 182(4112), 553–559. 10.1126/science.182.4112.553 [PubMed: 17739714]

Selnes OA, Jacobson L, Machado AM, Becker JT, Wesch J, Miller EN, & … McArthur JC (1991). Normative data for a brief neuropsychological screening battery. Multicenter AIDS cohort study. Perceptual and Motor Skills, 73(2), 539–550. 10.2466/pms.1991.73.2.539 [PubMed: 1766784]

Settles IH, O'Connor RC, & Yap SCY (2016). Climate perceptions and identity interference among undergraduate women in STEM: The protective role of gender identity. Psychology of Women Quarterly, 40(4), 488–503. 10.1177/0361684316655806

Shi D, DiStefano C, Maydeu-Olivares A, & Lee T (2022). Evaluating SEM model fit with small degrees of freedom. Multivariate Behavioral Research, 57(2–3), 179–207. 10.1080/00273171.2020.1868965 [PubMed: 33576257]

Suchy-Dicey AM, Eyituoyo H, O'Leary M, Cole SA, Traore A, Verney S, & … Whitney P (2022). Psychological and social support associations with mortality and cardiovascular disease in middle-aged American Indians: The Strong Heart Study. Social Psychiatry and Psychiatric Epidemiology, 57(7), 1421–1433. 10.1007/s00127-022-02237-7 [PubMed: 35157091]

Suchy-Dicey AM, Howard B, Longstreth WT Jr., Reiman EM, & Buchwald D (2022). APOE genotype, hippocampus, and cognitive markers of Alzheimer's disease in American Indians: Data from the Strong Heart Study. Alzheimers Dementia, 18, 2518–2526. 10.1002/alz.12573

Suchy-Dicey AM, Muller CJ, Madhyastha TM, Shibata D, Cole SA, Zhao J, & … Buchwald D (2018). Telomere length and magnetic resonance imaging findings of vascular brain injury and central brain atrophy: The Strong Heart Study. American Journal of Epidemiology, 187(6), 1231–1239. 10.1093/aje/kwx368 [PubMed: 29860472]

Suchy-Dicey AM, Muller CJ, Shibata D, Howard BV, Cole SA, Longstreth WT Jr., & … Buchwald D (2021). Comparing vascular brain injury and stroke by cranial magnetic resonance imaging, physician-adjudication, and self-report: Data from the Strong Heart Study. Neuroepidemiology, 55(5), 398–406. 10.1159/000517804 [PubMed: 34428763]

Suchy-Dicey AM, Oziel K, Sawyer C, Olufadi Y, Ali T, Fretts AM, Umans JG, Shibata DK, Longstreth WT, Rhoads K, & Buchwald DS, (2022). Educational and clinical associations with longitudinal cognitive function and brain imaging in American Indians: The Strong Heart Study. Neurology, 99(24), e2637–e2647. 10.1212/WNL.0000000000201261 [PubMed: 36289000]

Suchy-Dicey AM, Shibata D, Best LG, Verney SP, Longstreth WT Jr., Lee ET, & … 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. 10.1159/000443277 [PubMed: 27603047]

Suchy-Dicey AM, Shibata D, Cholerton B, Nelson L, Calhoun D, Ali T, & … Verney SP (2020). Cognitive correlates of MRI-defined cerebral vascular injury and atrophy in elderly American Indians: The Strong Heart Study. Journal of the International Neuropsychological Society, 26(3), 263–275. 10.1017/S1355617719001073 [PubMed: 31791442]

Suchy-Dicey AM, Verney SP, Nelson LA, Barbosa-Leiker C, Howard BA, Crane PK, & Buchwald DS (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. 10.1111/jgs.16434 [PubMed: 32250446]

Tombaugh TN, Kozak J, & Rees L (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. Archives of Clinical Neuropsychology, 14(2), 167–177. https://www.ncbi.nlm.nih.gov/pubmed/14590600 [PubMed: 14590600]

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

Veizel K, & Zibulsky J (2013). Wide range achievement test–fourth edition. In Reynolds CR, Vannest KJ, & Fletcher-Janzen E (Eds.), Encyclopedia of special education: A reference for the education of children, adolescents, and adults with disabilities and other exceptional individuals. Wiley.

Verney SP, Suchy-Dicey AM, Cholerton B, Calhoun D, Nelson L, Montine TJ, & … Buchwald D (2019). The associations among sociocultural factors and neuropsychological functioning in older American Indians: The Strong Heart Study. Neuropsychology, 33(8), 1078–1088. 10.1037/neu0000574 [PubMed: 31343235]

Widaman KF (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. Monographs of the Society for Research in Child Development, 71(3), 42–64. 10.1111/j.1540-5834.2006.00404.x

Zahodne LB, Manly JJ, Smith J, Seeman T, & Lachman ME (2017). Socioeconomic, health, and psychosocial mediators of racial disparities in cognition in early, middle, and late adulthood. Psychology and Aging, 32(2), 118–130. 10.1037/pag0000154 [PubMed: 28287782]

Zahodne LB, Sharifian N, Kraal AZ, Zaheed AB, Sol K, Morris EP, & … Brickman AM (2021). Socioeconomic and psychosocial mechanisms underlying racial/ethnic disparities in cognition among older adults. Neuropsychology, 35(3), 265–275. 10.1037/neu0000720 [PubMed: 33970660]

Zhang Y, Galloway JM, Welty TK, Wiebers DO, Whisnant JP, Devereux RB, & … Lee ET (2008). Incidence and risk factors for stroke in American Indians: The Strong Heart Study. Circulation, 118(15), 1577–1584. 10.1161/CIRCULATIONAHA.108.772285 [PubMed: 18809797]

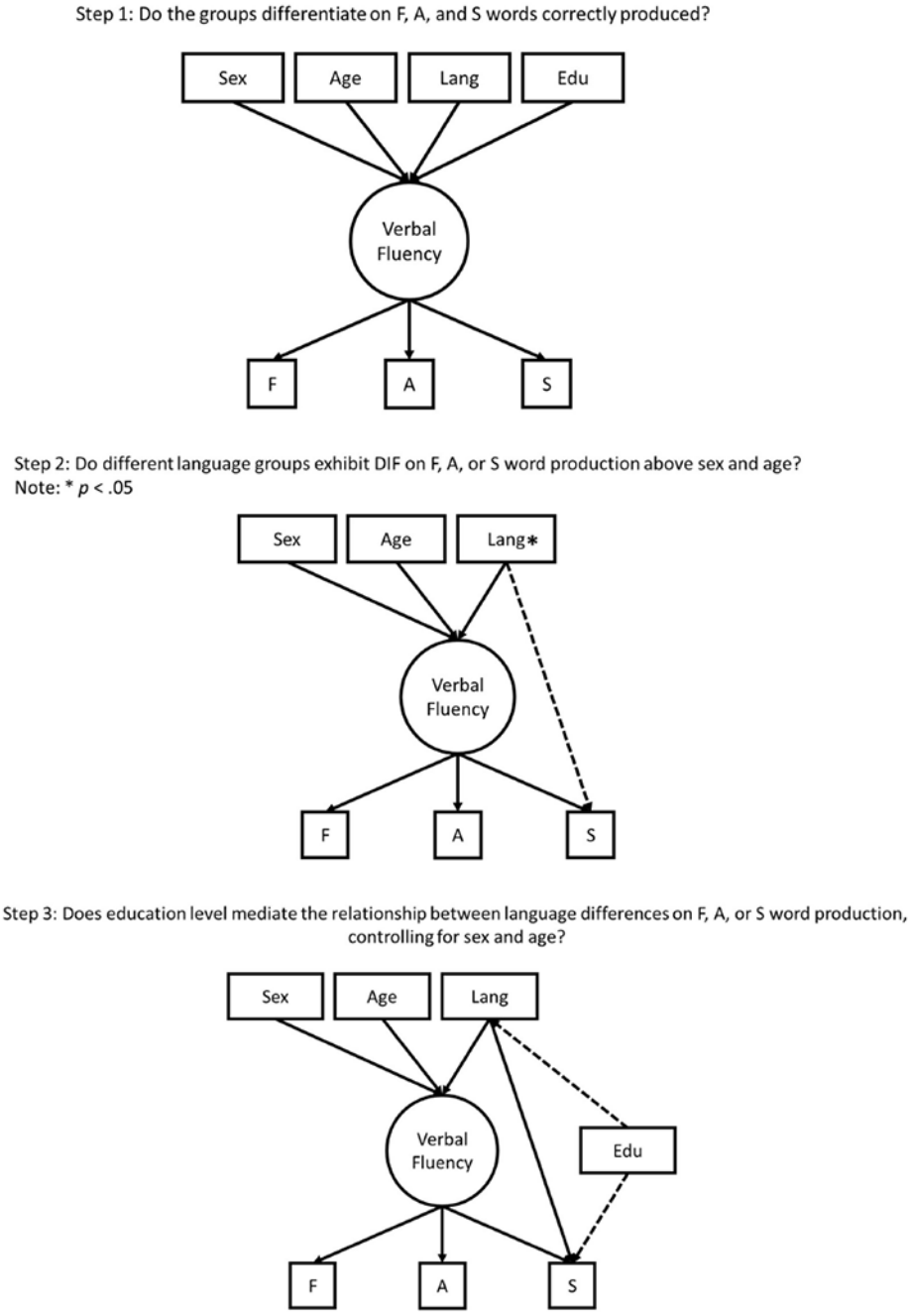

Figure 1.

A Schematic Example of the MIMIC Framework for Mediated DIF Models.

Note. MIMIC = multiple indicator multiple cause; DIF = differential indicator analysis.

Table 1.

Selected Participant Characteristics Among American Indian Participants.

Group	Visit 1 2010–2013	Visit 2 2017–2019
Total *N*	818	403
Age, *M* (*SD*)	73.0 (5.9)	78.0 (4.7)
Female, %	67.5	70.4
Years of education, *M* (*SD*)	12.6 (2.7)	13.0 (2.6)
Bilingual ability: moderately or very well, %	39.4	33.7
COWA score, *M* (*SD*)	24.4 (11.4)	24.3 (11.1)
WRAT score, *M* (*SD*)	—	40.6 (9.2)

Note. WRAT scores were not measured at Visit 1. WRAT = Wide Range Achievement Test.

Table 2.

Omega Internal Consistency Reliability Estimates in COWA Scores, by Study Strata.

Group	*N*	ω
Total	818	.88
Sex		
Male	264	.89
Female	549	.88
Age in years		
65–69	279	.84
70–79	413	.90
80 +	121	.89
Education in years		
⩽12	438	.88
>13	375	.86
Native language spoken well		
Not at all	260	.89
A little	233	.86
Moderately well	106	.88
Very well	214	.89

Note. Ω = omega internal consistency reliability estimate; COWA = Controlled Oral Word Association.

Table 3.

Test–Retest Correlations in COWA Scores, by Study Strata.

Group	*M* (*SD*) Visit 1	*M* (*SD*) Visit 2	Pearson's *r*
Total	24.4 (11.4)	24.3 (11.1)	.77
Sex			
Male	23.4 (11.5)	22.5 (11.3)	.75
Female	24.8 (11.4)	25.1 (11.0)	.77
Age in years			
60–69	26.9 (10.3)	26.5 (11.0)	.76
70–79	23.7 (11.7)	22.8 (11.1)	.76
80 +	20.7 (11.7)	21.5 (9.9)	.80
Education in years			
⩽12	20.5 (10.0)	21.1 (10.0)	.76
>13	28.9 (11.3)	27.4 (11.3)	.73
Native language spoken well			
Not at all	26.0 (11.5)	25.4 (11.6)	.79
A little	25.1 (10.9)	24.0 (11.6)	.75
Moderately well	27.2 (12.2)	24.3 (11.3)	.70
Very well	20.3 (10.3)	23.0 (9.4)	.80

Note. COWA = Controlled Oral Word Association.

All Pearson correlations were significant, $p < .05$; all pairwise comparisons of stability coefficients were insignificant, $p > .05$.

Table 4.

Standardized Parameter Estimates From the MIMIC Model, Visit 1 (N = 818, 2010–2013).

Group	Estimate	Standard error	*p* value	Cohen's *d*
Step 1: Differences in scores across groups				
Male Sex	−.093	.034	.006	0.201
Age	−.187	.034	<.001	0.488
Education	.364	.032	<.001	0.798
Language	−.073	.035	.034	0.151
Step 2: Language group differences across letter tasks				
Letter''F''	.028	.195	.885	—
Letter ''A''	−.120	.183	.511	—
Letter ''S''	.112	.212	.599	—

Note.''—'' indicates not applicable.

Table 5.

Standardized Parameter Estimates From the MIMIC Model, Visit 2 (N = 395, 2017–2019).

Group	Estimate	Standard error	*p* value	Cohen's *d*
Step 1a: Differences in scores across groups using education in years				
Male sex	−.069	.051	.176	0.019
Age	−.106	.052	.039	0.316
Education	.324	.048	<.001	0.096
Language	−.086	.052	.095	0.024
Step 1b: Differences in scores across groups using WRAT scores				
Male sex	−.060	.048	.212	0.016
Age	−.077	.049	.112	0.256
WRAT	.477	.042	<.001	1.24
Language	−.061	.049	.213	0.017
Step 2: Language group differences across letter tasks				
Letter ''F''	.005	.035	.881	—
Letter ''A''	−.005	.034	.873	—
Letter ''S''	.001	.035	.986	—

Note. ''—'' indicates not applicable. Step 2 was only conducted once across the two models because education and WRAT scores were not involved in assessing differential test effects. The effective sample size was 395, given that nine cases had missing values. WRAT = Wide Range Achievement Test.