Published in final edited form as:

Neuropsychology. 2011 March ; 25(2): 260–269. doi:10.1037/a0021090.

Measurement Invariance of Neuropsychological Tests in Diverse Older Persons

Dan Mungas¹, Keith F. Widaman², Bruce R. Reed^{1,3}, and Sarah Tomaszewski Farias¹
University of California, Davis

Abstract

Objective—Comparability of meaning of neuropsychological test results across ethnic, linguistic, and cultural groups is important for clinicians challenged with assessing increasing numbers of older ethnic minorities. We examined the dimensional structure of a neuropsychological test battery in linguistically and demographically diverse older adults.

Method—The Spanish and English Neuropsychological Assessment Scales (SENAS), developed to provide psychometrically sound measures of cognition for multi-ethnic and multilingual applications, was administered to a community dwelling sample of 760 Caucasians, 443 African Americans, 451 English-speaking Hispanics, and 882 Spanish-speaking Hispanics. Cognitive function spanned a broad range from normal to mildly impaired to demented. Multiple group confirmatory factor analysis (CFA) was used to examine equivalence of the dimensional structure for the SENAS across the groups defined by language and ethnicity.

Results—Covariance among 16 SENAS tests was best explained by five cognitive dimensions corresponding to episodic memory, semantic memory/language, spatial ability, attention/working memory, and verbal fluency. Multiple group CFA supported a common dimensional structure in the diverse groups. Measures of episodic memory showed the most compelling evidence of measurement equivalence across groups. Measurement equivalence was observed for most but not all measures of semantic memory/language and spatial ability. Measures of attention/working memory defined a common dimension in the different groups, but results suggest that scores are not strictly comparable across groups.

Conclusions—These results support the applicability of the SENAS for use with multi-ethnic and bilingual older adults, and more broadly, provide evidence of similar dimensions of cognition in the groups represented in the study.

Keywords

Neuropsychological Assessment; Measurement Equivalence; Ethnicity; English; Spanish

Neuropsychological assessment is a core component of the diagnostic process for Alzheimer's disease (AD) and other dementias, and plays a major role in monitoring

Correspondence to: Dan Mungas, Ph.D., Department of Neurology, 4860 Y Street, Suite 3700, Sacramento, CA 95817. Phone: (916)734-5730, FAX: (916)456-9350, dmmungas@ucdavis.edu.

¹Department of Neurology, School of Medicine, University of California, Davis

²Department of Psychology, University of California, Davis

³Veterans Administration Northern California Health Care System

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/neu

response to treatment. In the U.S., numbers of older adults of ethnic minority heritage are increasing in absolute terms and as a percentage of the total population. Although the general population over age 65 is projected to increase by 160% over the next 50 years, Hispanic older adults are projected to increase over 600% and African American older adults by 270% (Population Reference Bureau, 2000; U.S. Bureau of the Census and J.C. Day, 1993). As increasingly larger numbers of ethnic minorities seek evaluation and treatment of memory and other cognitive changes, comparability of meaning of neuropsychological test scores across diverse groups becomes particularly relevant.

Concern about measurement equivalence or invariance arose in reference to the fairness of college admission tests (Meredith & Teresi, 2006). At issue was whether tests were biased towards the selection of one sub-population at the expense of another; and this led to development of methods to examine measurement equivalence among groups. Valid inferences about individual differences in attributes/abilities of persons from different groups cannot be made unless the attributes relate to observed test results in the same way in the different groups (Borsboom, 2006). This condition, known as measurement equivalence, essentially means that two persons from different groups who have the same true ability will achieve the same test score, and is a prerequisite for unbiased measurement (Meredith & Teresi, 2006; Vandenberg & Lance, 2000). In a factor analytic context, measurement equivalence is discussed under the label of factorial invariance and is evaluated by examining the relations between theoretical (latent) constructs (i.e. memory) and their manifest (observed) indicators (test scores from measures of memory). Psychological measurements are considered comparable when relations between test scores and factors defining latent constructs are equivalent across groups (Reise, Widaman, & Pugh, 1993). Without equivalent measurement, observed scores from diverse groups are expressed in different scales and are not directly comparable (Drasgow & Kanfer, 1985). For example, a score of 15 might be associated with higher ability in one group than in another, and so, would not have the same meaning for individuals from these two groups.

Principal component analysis (PCA) was widely used in earlier studies to investigate measurement invariance; for example, to examine of the construct validity of intelligence tests across groups defined by race/ethnicity (e.g. Chastain & Joe, 1987; Kaufman, McLean, & Reynolds, 1991) and language (e.g. Drago, Daum, & Canavan, 1991; Gomez, Piedmont, & Fleming, 1992; Insua, 1983; Laicardi, Frustaci, & Lauriola, 1996; Orsini & Laicardi, 2000). PCA has limitations in evaluating measurement invariance (Borsboom, 2006), and more rigorous methods based on confirmatory factor analysis (CFA) have been recommended more recently for comparing the factor structure of cognitive tests and batteries across diverse groups (e.g. Geary & Whitworth, 1988; Grégoire, 2004; Pedraza, et al., 2005; Widaman & Reise, 1997). In CFA, the factor structure is specified *a priori* on the basis of previous research and/or theoretical considerations (McArdle, 1996). To test the dimensional equivalence across multiple, independent groups the *a priori* factor structure is fit to all groups simultaneously. Factor solutions shown to be identical across groups provide strong evidence of common dimensions underlying performance.

This study examined the dimensional structure of a neuropsychological test battery designed to measure neuropsychologically relevant domains of cognition in linguistically and demographically diverse older adults. The Spanish and English Neuropsychological Assessment Scales (SENAS) has undergone extensive development as a battery of cognitive tests relevant to diseases of aging (Mungas, Reed, Crane, Haan, & González, 2004; Mungas, Reed, Haan, & Gonzalez, 2005; Mungas, Reed, Marshall, & González, 2000; Mungas, Reed, Tomaszewski Farias, & DeCarli, 2005). Here the dimensional structure of the SENAS was evaluated in groups of older African Americans, Caucasians, and English- and Spanish-speaking Hispanics. This study represents one component step in the ongoing construct

validation of the SENAS and also addresses more general questions about the organization of cognition in older populations.

Methods

Participants

Participants were community-dwelling Caucasian, African American, and Hispanic individuals aged 60 or older who spoke either English or Spanish and were recruited between 1998 and 2008 as part of the ongoing development of the SENAS (Mungas, et al., 2004; Mungas, et al., 2000). A variety of recruitment methods, designed to identify and recruit individuals with cognitive functioning representative of the community dwelling population, were used and are described in detail elsewhere (Hinton, et al., In Press; Mungas, et al., 2004). Ninety-three percent of participants were recruited using community survey methods; the remainder were patients seeking clinical services related to cognitive decline. All participants signed informed consent under protocols approved by institutional review boards at University of California at Davis, the Veterans Administration Northern California Health Care System, and San Joaquin General Hospital in Stockton, California.

All clinic patients and a stratified subsample of community participants received a comprehensive, multidisciplinary clinical evaluation that included a detailed medical history, physical exam, and neurological exam. This evaluation also included a clinical neuropsychological evaluation using standard neuropsychological tests that are not part of the SENAS. Routine dementia work-up laboratory tests were obtained. A bilingual physician examined Spanish-speaking patients. A family member or other informant with close contact with the participant was interviewed to obtain information about level of independent functioning. Diagnosis of cognitive syndrome (Normal Cognition, Mild Cognitive Impairment (MCI), Dementia) and, in the instance of dementia, underlying etiology was made according to standardized criteria and methods (DeCarli, et al., 2008; Haan, et al., 2003).

Diagnosis was not used in analyses for this study but was used to describe the clinical characteristics of this sample. Sampling percentages were used as weights to relate the community participant subsample that had received clinical evaluation back to the overall community sample and to estimate the prevalence of specific diagnostic categories in the community sample. Dementia prevalence was 9.4% in African Americans, 11.2% in Caucasians, 14.9% in Hispanics tested in English, and 15.2% in Hispanics tested in Spanish (African Americans and Hispanic-Spanish significantly differed using a Bonferroni corrected p value of .05/6=.008). Estimated MCI prevalence was 29.3% in African Americans, 22.7% in Caucasians, 14.5% in Hispanics tested in English, and 27.1% in Hispanics tested in Spanish (Hispanic-English prevalence was significantly lower than for the other three groups and Caucasians had significantly lower prevalence of MCI than African Americans). Normal Cognition prevalence was 61.2% in African Americans, 66.1% in Caucasians, 70.5% in Hispanics tested in English, and 57.7% in Hispanics tested in Spanish (Hispanic-Spanish was significantly lower than Hispanic-English and Caucasian, African American was lower than Hispanic-English). While there were some minor groups differences, the overall distributions of clinical diagnoses related to cognitive impairment were similar, and more importantly, there was a range from demented to fully normal in all groups.

Demographic characteristics are summarized in Table 1. The overall sample included 760 Caucasian, 443 African American, 451 English-speaking Hispanic, and 882 Spanish-speaking Hispanic older persons. Hispanic participants primarily were of Mexican heritage. African American and Caucasian participants were tested in English.

Language of test administration was determined by an algorithm that utilized information about the participant's preferred language, utilization of English and Spanish, and which language was used in several specific contexts: converse in the home, listen to radio or TV, converse with family or friends outside the home, converse with others outside the home, and read (Mungas, et al., 2004). Eighty-two percent of Hispanics tested in English were bilingual -that is spoke, both English and Spanish well or very well. In contrast, only 16% of Hispanics tested in Spanish were bilingual. Preferred language was the language of administration in 96% of the Hispanic participants.

SENAS Measures

The SENAS has undergone extensive development as a battery of cognitive tests relevant to diseases of aging (Crane, et al., 2008; Mungas, et al., 2004; Mungas, Reed, Haan, et al., 2005; Mungas, Reed, Tomaszewski Farias, et al., 2005; Mungas, et al., 2000). Modern psychometric methods based on item response theory were used to create psychometrically matched measures across different scales and across English and Spanish versions. The Supplement contains more detailed information about the cognitive domains assessed by the SENAS and the specific measures of those domains. A set of 13 measures in domains other than executive function were part of the original version of the SENAS; psychometric properties are described in Mungas et al. (2004). The executive function tasks are newer additions to the SENAS. Three composite scores measuring Category Fluency, Phonemic (letter) Fluency, and Working Memory (digit span backward, visual span backward, list sorting) were created using item response theory methods previously reported (Crane, et al., 2008; Mungas, Reed, Tomaszewski Farias, et al., 2005). A total of 16 SENAS variables were used in subsequent factorial invariance analyses.

Data Analysis

Overview—Confirmatory factor analysis (CFA) was used to evaluate the dimensional structure of the 16 SENAS measures and to test measurement invariance across the groups defined by ethnicity and language. We first identified a best model to account for the dimensional structure of the SENAS measures in each of the four ethnic/language groups. This was done separately for each group, and CFA was used to compare fit of multiple, competing, hypothesized dimensional structures. The second step of analyses utilized multiple group CFA to evaluate invariance of model parameters across the four ethnic/language groups.

Basic dimensional structure—The 16 SENAS variables were modeled as observed indicators of underlying latent dimensions. Eight competing models to account for covariance among these SENAS measures were evaluated (Table 2). Three four-factor models were evaluated that differed with respect to the two fluency measures; the fluency variables were indicators of executive function in one model, language in a second, and both executive function and language in a third. Independent analyses were performed within each ethnic/language group to evaluate these competing models and arrive at a best overall model within each group. Indicators initially were allowed to load on only one dimension. However, cognitive scores often are complexly determined by more than one latent dimension, and a subsequent process identified indicators related to more than one latent dimension. This was done with a specification-search approach (Jöreskog & Sörbom, 1996) guided by conceptual considerations.

Invariance of best model—Multiple group CFA was used to test invariance of the best model across the four ethnic/language groups. Multiple group CFA uses formal, hypothesis driven tests to determine whether model parameters are invariant across groups. Generally, means and variances of latent dimensions are allowed to differ across groups, reflecting real-

world likelihoods that underlying distributions of abilities are not identical. All other model parameters are initially constrained to be the same in the different groups. Then, a systematic process is followed in which different parameters are allowed to have different values across groups, and related improvement in overall model fit is tested. Previous literature in this area has identified different levels of invariance that have different interpretive significance (e.g. Gregorich, 2006; Meredith & Teresi, 2006; Widaman & Reise, 1997). Configural invariance occurs when the same dimensional structure is applicable in the different groups, though formal parameters like loadings (the strength of the relationship of the test with the underlying factor) and intercepts (the expected test score when the factor score is 0.0, i.e. at the sample average) might be different. Weak factorial invariance adds an additional criterion that factor loadings do not differ across groups. This means that observed variables relate to the latent factors in the same way. Strong factorial invariance is more restrictive, requires that intercepts also are equated, and means that the absolute scores of the observed variables can be directly compared across groups. Strict invariance is the most restrictive standard, requiring the same loadings, intercepts, and residual variances (the variation in the test score not explained by the factor). When present, individual differences in test scores are completely determined by underlying differences in latent abilities and scores are strictly comparable across groups. Partial invariance occurs when some parameters differ across groups, but many are invariant.

Model estimation was performed with Mplus version 5.2 (Muthén & Muthén, 1998–2008) using a maximum likelihood estimator for continuous variables applied to a mean and covariance data structure. Latent variable modeling traditionally uses an overall chi square test of model fit, often supplemented by a number of fit indices to better characterize model fit. Commonly used fit indices include the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the root mean square error of approximation (RMSEA; Browne & Cudek, 1993), and the standardized root mean squared residual (SRMR; Bentler, 1995). The chi-square difference test (Steiger, Shapiro, & Browne, 1985) was used to determine if fit significantly improved as a result of freeing one or more parameters in a model.

Technical details about methods for data analysis are presented in the Supplement.

Results

Demographic Characteristics

Mean education markedly differed across ethnic groups ($F[3,2536]=954.7, p<.001$). Caucasians had about 14 years of education on average, African Americans 13 years, Hispanics tested in English 11 years, and Hispanics tested in Spanish less than 5 years. All pairwise comparisons of means were statistically significant using a Bonferroni adjusted p value (.05/6=.008). There was a broad range of education in all groups. Age differences were significant but less substantial ($F[3,2536]=35.2, p<.001$); Hispanics were two to three years younger than Caucasians and African Americans, but the two Hispanic groups did not differ significantly nor did Caucasians and African Americans. Group differences in gender were also significant ($\chi^2[3,2536]=27.2, p=.001$); percentage of females was highest in African Americans, lowest in Caucasians, and intermediate in Hispanics.

Dimensionality

The first step in model building was to identify the model that best represented the cognitive dimensions underlying the 16 SENAS measures. The eight alternative models presented in Table 2 were evaluated independently in each of the four groups, yielding highly similar results across groups. The one- and two-factor models provided the poorest fit to the data.

The four-factor models provided relatively good and similar fit across groups, but the five factor model provided better fit. The six and seven factor models had the best fit, but were problematic because the Semantic, Verbal, and Abstraction factors were too highly correlated to comprise distinct latent dimensions. Consequently, the five factor model was chosen as the best model for further analyses.

Three changes to the five factor model empirically improved model fit in all four groups and were conceptually supported. First, residual correlation was allowed between Object Naming and Picture Association. These two scales have consistently been very highly correlated in previous development work with the SENAS and have defined a semantic memory factor (Mungas, et al., 2004), represented in this study in the six and seven factor models. Second, Spatial Configuration Learning was allowed to have non-zero loadings on both the Episodic Memory and Spatial factors. This complex structure was supported both by previous work with the SENAS (Mungas, et al., 2004) as well as by other studies showing that visual memory has both episodic memory and spatial components (Larrabee, Kane, & Schuck, 1983; Pedraza, et al., 2005; Smith, et al., 1992; Smith, Ivnik, Malec, & Tangalos, 1993). The final change to the five factor model allowed loadings for Verbal Comprehension on both the Semantic/Language and Attention/Working Memory factors. Verbal Comprehension involves language decoding of increasingly complex and longer auditory commands, and conceptually, has both language and working memory components. Model fit after incorporating these three modifications was good in all four groups (African American: $\chi^2[91,443]=171.6$, $p < 0.001$; CFI=0.975, TLI=0.968, RMSEA=0.045 (95% CI=0.034–0.055, SRMR=.032; Caucasian: $\chi^2[91,760]=296.1$, $p < 0.001$; CFI=0.958, TLI=0.944, RMSEA=0.055 (95% CI=0.048–0.062, SRMR=.052; Hispanic - English: $\chi^2[91,451]=161.3$, $p < 0.001$; CFI=0.974, TLI=0.965, RMSEA=0.041 (95% CI=0.031–0.052, SRMR=.039; Hispanic - Spanish: $\chi^2[91,882]=255.2$, $p < 0.001$; CFI=0.974, TLI=0.965, RMSEA=0.045 (95% CI=0.039–0.052, SRMR=.031). Factor correlations from this best model in the ethnic/language groups generally were in the 0.70–0.90 range. Correlations of Episodic Memory with Semantic/Language, Spatial, and Attention/Working memory were generally in the 0.65–0.75 range. The strongest intercorrelations of factors involved Semantic/Language and Spatial, with values in the four groups ranging from 0.87–0.95.

Measurement Invariance across Ethnic/Language Groups

In the next step, multiple group CFA was used in a series of analyses to evaluate factorial invariance across groups. Multiple group CFA simultaneously tests whether a hypothesized factor structure fits equally well in different groups. The first model tested the hypothesis of common dimensional structure (configural invariance). The modified five-factor model from the previous analyses was specified for all four groups, but loadings, intercepts, and residual variances were allowed to differ across groups (Model 1). This model fit well ($\chi^2[367,2536]=888.8$, $p < 0.001$; CFI=0.969, TLI=0.960, RMSEA=0.047 (95% CI=0.043–0.051), SRMR=0.039), indicating that the five factor dimensional structure was shared across groups.

We then tested equivalence of loadings by constraining loadings to be equal across groups (i.e. requiring all tests to be equally related to the factor across the groups). Overall model fit was still good ($\chi^2[406,2536]=1248.1$, $p < 0.001$; CFI=0.951, TLI=0.942, RMSEA=0.057 (0.054–0.061), SRMR=0.062) but fit was significantly poorer than in the less constrained configural invariance model, and six (out of 72) loadings were identified that did significantly differ across groups. (See Supplement for additional detail.) Model fit was good when these six additional parameters were allowed to differ (Model 2; $\chi^2[400,2536]=957.8$, $p < 0.001$; CFI=0.967, TLI=0.961, RMSEA=0.047 (0.043–0.051), SRMR=0.043).

The next analysis started with Model 2 and constrained all intercepts to be equal (i.e., requiring that a factor score of 0.0 would correspond to the same test score in different groups). This resulted in significantly poorer model fit in comparison with Model 2 ($\chi^2[33,2536]=944.1$, $p < .001$) and relatively poor absolute model fit ($\chi^2[433,2536]=1901.9$, $p < 0.001$; CFI=0.914, TLI=0.904, RMSEA=0.073 (0.070–0.077), SRMR=0.093). We identified 13 out of 64 intercepts that differed across groups, and freely estimating these parameters resulted in a well fitting model (Model 3; $\chi^2[420,2536]=1014.1$, $p < 0.001$; CFI=0.965, TLI=0.960, RMSEA=0.047 (0.044–0.051), SRMR=0.047).

The next step started with Model 3 and constrained residual variances to equality (i.e. the amount of variance in the tests explained by the factors was equal across groups). This resulted in significantly poorer model fit in comparison with Model 3 ($\chi^2[48,2536]=422.9$, $p < .001$) and poorer but acceptable absolute model fit ($\chi^2[468,2536]=1437.0$, $p < 0.001$; CFI=0.943, TLI=0.942, RMSEA=0.057 (0.054–0.061), SRMR=0.075). We found 10 out of 64 residual variances that differed across groups, and freely estimating these parameters resulted in a well fitting model (Model 4; $\chi^2[458,2536]=1084.0$, $p < 0.001$; CFI=0.963, TLI=0.961, RMSEA=0.047 (0.043–0.050), SRMR=0.049).

A second iteration started with Model 4 and systematically tested for additional non-invariant loadings, then intercepts, and residual variances. One additional non-invariant loading and one non-invariant intercept were identified in this process, and freely estimating these parameters along with the other previously identified non-invariant parameters resulted in good model fit.

Table 3 presents standardized factor loadings from the final model that identified and freely estimated non-invariant parameters. Intercepts and residual variances from this final model are included in the Supplement. Standardized loadings can be interpreted as correlations of the observed variables with the latent factors. While several parameters significantly differed across groups, the sample size for this study was relatively large, so power for detecting small differences was high. Consequently, many of the differences across groups were small. Clear differences in model parameters were noted for four of the 16 SENAS test scores. Verbal Expression showed striking differences in residual variances (about a fourfold difference between the smallest and largest), and in addition, showed a large intercept difference for African Americans in comparison with the other groups. Verbal Attention showed robust differences in intercepts. Any given Attention/Working Memory factor score was associated with highest Verbal Attention performance in African Americans and Caucasians, lowest performance in Spanish-speaking Hispanics, and intermediate performance in English-speaking Hispanics. The third variable, Verbal Conceptual Thinking, had a clearly different intercept in Spanish-speaking Hispanics. Visual Attention had moderate intercept differences for Hispanics in comparison with African Americans and Caucasians. A model that accounted for these difference in Verbal Expression, Verbal Attention, Verbal Conceptual Thinking, and Visual Attention, but that constrained residual variances, intercepts, and loadings to equality for the other 12 variables fit well ($\chi^2[471,2536]=1373.3$, $p < .001$; CFI=.947, TLI=.946, RMSEA=.055 (.052–.058), SRMR=.060).

In summary, four observed variables accounted for most of the measurement differences across groups. After accounting for these differences, good support was found for measurement invariance of the factor model involving the remaining 12 observed variables. A cross validation was performed by randomly dividing the overall sample into two equally sized subsamples, following the model specification process in each subsample, and validating the model identified in one subsample using the other independent subsample.

Results provided strong evidence that the approach for evaluating measurement invariance yielded stable and replicable results. (See Supplement).

Discussion

The population of older persons is growing rapidly and is becoming increasingly diverse. Neuropsychological assessment is important in older persons because it is central to identifying, monitoring, and diagnosing neurodegenerative disease of aging. Identifying useful dimensions of cognition and understanding the similarities and differences of these dimensions across groups defined by race, ethnicity, language, and culture has important implications for clinical and scientific understanding of both normal and abnormal cognition in late life.

This study found five dimensions that reasonably explained covariance among the 16 component measures from the SENAS in four different groups: African Americans, Caucasians, and English- and Spanish-speaking Hispanics. This five-factor model included latent variables representing Episodic Memory, Semantic Memory/Language, Spatial ability, Attention/Working Memory, and Fluency. It is axiomatic that factor structure is dependent on the variables that are entered into the analysis, and the SENAS is a specific battery of tests. This raises an important question of how generalizable these results are to the broader area of neuropsychological assessment of older persons. Consequently, the comparability of the dimensions identified in this study with previous literature on dimensionality of neuropsychological tests has relevance for understanding both the SENAS and the broader organization of cognition.

Overall, the results of this study are strikingly consistent with other studies using different test batteries applied to samples from different populations. The dimensions underlying the Wechsler Adult Intelligence Scale and Wechsler Memory Scale have been most often studied, and factors representing verbal abilities, visual perceptual abilities, attention/concentration, and memory, sometimes involving different sub components, have frequently been found (Bowden, Carstairs, & Shores, 1999; Bowden, Cook, Bardenhagen, Shores, & Carstairs, 2004; Larrabee, et al., 1983; Smith, et al., 1992; Smith, et al., 1993). These dimensions are well represented in this study. The Semantic/Language factor closely corresponds to the verbal abilities factor from previous studies, Spatial corresponds to visual perceptual abilities, Attention/Working Memory to attention/concentration, and Episodic Memory to memory. Further, visual memory measures previously have been found to load as strongly on a visual spatial ability factor as on a memory factor (Larrabee, et al., 1983; Pedraza, et al., 2005; Smith, et al., 1992; Smith, et al., 1993), and this is consistent with the results of this study. The Fluency factor in this study is somewhat novel, but fluency measures often have not been included in previous studies.

Attempts to identify dimensions of neuropsychological test performance in different linguistic, ethnic, or cultural groups have been limited. Pontón et al. (2000) performed a factor analytic study of the Neuropsychological Screening Battery for Hispanics (NeSBHIS) using a sample of Hispanics, and concluded that the results were consistent with previous neuropsychological literature, and thus supported the construct validity of the NeSBHIS. Because no direct comparison of results with a non-Hispanic sample was made, the Pontón et al. study could not address the question of measurement invariance. Measurement invariance of the Mayo Older Adult Normative Studies (MOANS) core battery across Caucasian and African American samples was tested in a limited manner. Smith and colleagues (Smith, et al., 1992; Smith, et al., 1993) used confirmatory factor analysis to initially examine the dimensional structure of the MOANS core battery in the MOANS normative sample and in a sample of mixed clinical patients, both comprised primarily of

Caucasians. Later, Pedraza et al. (2005) used confirmatory factor analysis to evaluate the applicability of the five factor solution identified in both Smith studies (Smith, et al., 1992; Smith, et al., 1993) to the MOANS core battery administered in an independent sample of older African Americans. While a five factor model based on the previous studies fit well in the African American sample, invariance of the measurement model was not directly tested across the three studies. Thus, the Pedraza et al. study provided support for generality of dimensions, but could not address more stringent and specific forms of measurement invariance.

Bowden et al. (2004) directly tested measurement invariance across clinic and community samples of a neuropsychological test battery primarily composed of the Wechsler scales. Bowden also examined measurement invariance across alcohol dependent and normal community groups (Bowden, et al., 2001). These studies used CFA methods to more systematically characterize measurement equivalence. Similar methods were applied in the area of aging and cognition by Tuokko et al. (2009) to evaluate measurement invariance across English and French language groups tested as part of the Canadian Study on Health and Aging (CSHA). They showed evidence of partial measurement invariance in the CSHA neuropsychological test battery and addressed implications about when comparisons across groups involving observed variables and derived latent variables are appropriate and meaningful.

The current study was unique in that it formally evaluated measurement invariance across multiple different ethnicity/language groups. Although some measurement differences across groups were found, results showed that the same dimensions were present in the four groups and that most SENAS subtests related to these dimensions in the same way across groups. Relatively few parameters were found to differ across groups, and, in many cases, group differences for these non-invariant parameters were small. In general, few group differences in loadings were found, and loadings were of similar magnitude across groups even when significant differences were found. This indicates that the same dimensions are being measured and that ordering of individuals on these dimensions within groups is being done in the same way. As a result, correlations of external variables with these dimensions and the specific tests that define them can be meaningfully compared across groups. For example, it would be meaningful to compare correlations of age with test scores across groups. Clinically, equivalent loadings mean that test scores and scores on latent factors derived from them can be used in the same way within each of the 4 groups studied here. For example, if the SENAS is used to measure individual differences within groups or to measure longitudinal change of individuals, the results could be similarly interpreted regardless of ethnicity/language.

Differences involving intercepts were more substantial, both in number of variables involved and in magnitude of observed differences. Where intercept differences are present, this indicates that the absolute levels of performance may not be comparable across groups, even if individual differences within groups have the same meaning. That is, score of 10 might have a different meaning in two different groups, even if a change from a 10 to an 8 would have the same meaning. A clinical implication is that group specific normative means are needed for individual tests, and possibly for factor scores. Differences in residual variances similarly point to the need for different normative standard deviations across groups.

Results of this study have practical significance for the use of the SENAS specifically and neuropsychological tests more broadly for assessing cognition in diverse older persons. For the Episodic Memory factor, only one parameter differed across groups, and the magnitude of group differences for this parameter was negligible. This indicates that the latent factor

and the observed variables that define it essentially meet requirements for strict factorial invariance. All three observed measures defining this factor provide equivalent measurement of episodic memory across groups. Group-specific norms are not necessary for Episodic Memory and its defining scales. Results were similar for the Fluency factor, though there were minor differences in Phonemic Fluency intercepts suggesting that this task is harder for Hispanics, especially Spanish speakers, in the sense that test scores for a given level of ability would be expected to be somewhat higher (about 0.25–0.33 s.d.) in African Americans and Caucasians. Group specific norms might be appropriate for this measure.

Results for the Semantic/Language, Spatial, and Attention/Working Memory were more complex. Some loadings differed across groups for Semantic/Language (3 of 20), Spatial (1 of 16), and Attention/Working memory (2 of 16), but differences in loadings were small and there were sufficient numbers of invariant loadings in each group to establish common latent dimensions with the same meanings in the different groups.

Five intercepts out of 20 differed across groups for Semantic/Language, but each group had at least two invariant intercepts. This is sufficient to establish a common metric that makes it appropriate to compare factor means across groups; comparison of means of individual tests defining this factor is appropriate except that the Verbal Expression score in African Americans and the Verbal Conceptual Thinking score in Spanish administration Hispanics likely are not on the same scale as these scores in other groups. Two of 16 intercepts for Spatial differed across groups and three indicators had either all invariant intercepts (Pattern Recognition, Non Verbal Memory) or minor non-invariant intercepts for one group (Spatial Localization). This indicates that group means of factor scores can be meaningfully compared and that group differences in Non Verbal Memory, Pattern Recognition, and Spatial Localization are meaningful. Intercept differences for Attention/Working Memory were more problematic; seven of 16 intercepts differed across groups, and magnitudes of differences were relatively large for Verbal Attention and Visual Attention. Only one indicator (Working Memory) had invariant intercepts. These results argue for caution in interpreting mean differences or in comparing scores across groups, especially when comparing Hispanics with Caucasians and African Americans.

The strongest evidence of measurement differences involved the intercept parameters for the Verbal Attention observed variable, a digit span task. Results indicated that, for a given level of latent Attention/Working Memory ability, expected performance of Caucasians and African Americans would be best, it would be worst for Spanish speaking Hispanics, and intermediate for English-speaking Hispanics. This pattern of results is quite consistent with a relatively large body of research examining digit span in different linguistic groups. Studies have shown cross-language differences in digit span, with better average performance in Chinese, intermediate performance in English, and poorer performance in Welsh and Spanish (Chen, Cowell, Varley, & Wang, 2009; Ellis & Hennelly, 1980; Olazaran, Jacobs, & Stern, 1996; Stigler, Lee, & Stevenson, 1986). Researchers have argued that these language differences are related to different phoneme lengths and articulation times of digits in the different languages (Chen, et al., 2009; Chincotta & Underwood, 1997; Murray & Jones, 2002; Olazaran, et al., 1996). There is evidence of a fixed temporal span of immediate attention; increased phoneme length results in longer presentation and rehearsal time so that less information is encoded within the fixed temporal span resulting in reduced retrieval of the information that was presented (Chen & Stevenson, 1988; Stigler, et al., 1986). Applied to this study, if the actual immediate attention span is the same in the different groups, the Spanish administration group would be expected to have the lowest digit span because Spanish digits have longer phonemes, and consequently more information must be processed and encoded for a given number of digits. The expected Verbal Attention Span performance of English-speaking Hispanics was greater than that for

Spanish-speaking Hispanics, but was less than for African Americans and Caucasians. Many of the English-speaking Hispanic group were bilingual. Digit span differences in bilinguals have been previously reported and attributed to testing in a non-native language (da Costa Pinto, 1991) or to reduced fluency in the language of testing (Chincotta & Hoosain, 1995).

There was a related finding involving the Attention/Working Memory dimension; both Hispanic groups had better expected performance on Visual Attention for the same underlying ability. This finding is suggestive of a relative advantage in visual attention span for Hispanics in comparison with the other groups. The reason for this advantage is not entirely clear, and previous studies have not shown differences in spatial span related to Spanish versus English test administration (Gasquoine, Croyle, Cavazos-Gonzalez, & Sandoval, 2007; Olazaran, et al., 1996). For reasons previously discussed, group mean differences related to the Attention/Working Memory factor should be considered with caution.

In summary, results of this study showed that Episodic Memory and Fluency provide equivalent measurement across groups. Semantic/Language and Spatial are appropriate for comparison of correlations of factors with external variables, and factor scores and some individual indicators can be meaningfully compared across groups. Correlations of Attention/Working Memory with external variables are supported, but cross-group comparisons of factor scores and observed indicators are more problematic. The individual SENAS variables that were most problematic in terms of measurement equivalence were Verbal Attention, Verbal Expression, Visual Attention, and Verbal Conceptual Thinking. Reference to group-appropriate norms would be important for interpreting these tests.

The use of shared variance techniques like factor analysis in evaluating the construct validity of neuropsychological tests has been controversial. Factor analysis has played a major role in construct validity studies in neuropsychology for many years and continues to be commonly used. Delis et al. (2003) challenged this practice and specifically questioned the appropriateness of these methods applied to normative samples and mixed samples of clinical patients. They presented data showing different correlations of immediate and delayed recall measures from a verbal learning task in different homogeneous patient groups, and also showed different principal components analysis (PCA) factor structures of measures from the verbal learning task in different homogeneous groups. They argued that conceptually and clinically relevant dissociations of constructs like immediate and delayed recall are obscured by covariance analysis methods, and concluded that direct comparisons of conceptually different patient groups are ultimately required to evaluate construct validity. Bowden (2004) addressed methodological problems with the analyses Delis et al used to support their argument. He argued that multiple group covariance structure analyses based on mean and covariance data structures and not on simple correlations or PCAs are the appropriate analyses, and comparing individual correlations or comparing results of independent PCAs can be misleading. The issues raised by Delis et al and Bowden have implications for the interpretation of this study.

While this study used modern covariance structure analysis methods recommended by Bowden (2004), any factor analytic study is limited both by the variables that are included and by the participant sample and the sources and amount of variability within that sample. A specific set of neuropsychological test variables was used in this study, but results were consistent with a number of other studies using entirely different tests. The sample in this study had very broad heterogeneity of demographic characteristics, and in addition, there was broad variability of cognitive test performance and presumably cognitive impairment independent of demographic effects. However, while cognitive impairment and dementia were clearly represented in the sample, this study was not designed to directly compare

homogeneous patient groups with theoretically relevant, specific etiologies for dementia and cognitive impairment, for example, Alzheimer's disease and frontotemporal dementia. Validity of any test is purpose specific, and construct validity in particular is complexly defined by a pattern of empirical results addressing theoretically relevant questions about the tests and the constructs they are designed to measure. The dimensions identified in this study are applicable for characterizing a very broad range of cognitive ability. This should support assessment of cognitive function and change in cognitive function in relation to external variables including demographic characteristics, measures of brain structure and function, and presence and severity of diseases of aging. For many clinical applications, identifying and monitoring cognitive impairment is the most important application of neuropsychological assessment. The construct validity of the dimensions identified in this study with respect to differential diagnosis of specific diseases of aging and of focal brain lesions requires further study with samples specifically designed for those purposes. This study shows that dimensions of cognition are similar in demographically diverse groups, and this represents an important incremental step in the process of the construct validation of the SENAS test battery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by National Institute on Aging Grants AG10220 and AG10129 and by the California Department of Public Health Alzheimer's Disease Program Contracts 09-11406 and 09-11407.

References

- Borsboom D. The attack of the psychometricians. *Psychometrika*. 2006;71(3):425–440. [PubMed: 19946599]
- Bowden SC. The role of factor analysis in construct validity: Is it a myth? *Journal of the International Neuropsychological Society*. 2004;10(7):1018–1019. [PubMed: 15803564]
- Bowden SC, Carstairs JR, Shores EA. Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale-Revised and Wechsler Memory Scale-Revised scores in a healthy community sample. *Psychological Assessment*. 1999;11(3):339–344.
- Bowden SC, Cook MJ, Bardenhagen FJ, Shores EA, Carstairs JR. Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence*. 2004;32(4):363–389.
- Bowden SC, Ritter AJ, Carstairs JR, Shores EA, Pead J, Greeley JD, et al. Factorial invariance for combined Wechsler Adult Intelligence Scale*€ Revised and Wechsler Memory Scale*€ Revised scores in a sample of clients with alcohol dependency. *Clinical Neuropsychologist*. 2001;15(1):69–80. [PubMed: 11778580]
- Busse A, Bischof J, Riedel-Heller SG, Angermeyer MC. Subclassifications for mild cognitive impairment: prevalence and predictive validity. *Psychol Med*. 2003;33(6):1029–1038. [PubMed: 12946087]
- Chastain RL, Joe GW. Multidimensional relations between intellectual abilities and demographic variables. [Peer Reviewed]. *Journal of Educational Psychology*. 1987;79(3):323–325.
- Chen C, Stevenson HW. Cross-linguistic differences in digit span of preschool children. *Journal of Experimental Child Psychology*. 1988;46(1):150–158.
- Chen ZY, Cowell PE, Varley R, Wang YC. A cross-language study of verbal and visuospatial working memory span. *J Clin Exp Neuropsychol*. 2009;31(4):385–391. [PubMed: 18720176]
- Chincotta D, Hoosain R. Reading rate, articulatory suppression and bilingual digit span. *European Journal of Cognitive Psychology*. 1995;7(2):201–211.

- Chincotta D, Underwood G. Digit span and articulatory suppression: A cross-linguistic comparison. *European Journal of Cognitive Psychology*. 1997;9(1):89–96.
- Crane PK, Narasimhalu K, Gibbons LE, Pedraza O, Mehta KM, Tang Y, et al. Composite scores for executive function items: demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*. 2008;14(5):746–759. [PubMed: 18764970]
- da Costa Pinto A. Reading rates and digit span in bilinguals: The superiority of mother tongue. *International Journal of Psychology*. 1991;26(4):471–483.
- DeCarli C, Reed BR, Jagust W, Martinez O, Ortega M, Mungas D. Brain behavior relationships among African Americans, whites, and Hispanics. *Alzheimer's Disease and Associated Disorders*. 2008;22(4):382–391.
- Delis DC, Jacobson M, Bondi MW, Hamilton JM, Salmon DP. The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society*. 2003;9:936–946. [PubMed: 14632252]
- Drago MC, Daum I, Canavan AG. A factor-analytic study of the Wechsler Intelligence Scales as administered to a clinical sample in Chile. [Peer Reviewed]. *Personality and Individual Differences*. 1991;12(10):993–998.
- Drasgow F, Kanfer R. Equivalence of psychological measurement in heterogeneous populations. [Peer Reviewed]. *Journal of Applied Psychology*. 1985;70(4):662–680. [PubMed: 4086417]
- Ellis NC, Hennelly RA. A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*. 1980;71(1):43–51.
- Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, et al. Global prevalence of dementia: a Delphi consensus study. *Lancet*. 2005;366(9503):2112–2117. [PubMed: 16360788]
- Gasquoin PG, Croyle KL, Cavazos-Gonzalez C, Sandoval O. Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. *Archives of Clinical Neuropsychology*. 2007;22(8):991–1001. [PubMed: 17900857]
- Geary DC, Whitworth RH. Dimensional structure of the WAIS-R: A simultaneous multi-sample analysis. [Peer Reviewed]. *Educational and Psychological Measurement*. 1988;48(4):945–956.
- Gomez FC, Piedmont RL, Fleming MZ. Factor analysis of the Spanish version of the WAIS: The Escala de Inteligencia Wechsler para Adultos (EIWA). [Peer Reviewed]. *Psychological Assessment*. 1992;4(3):317–321.
- Grégoire J. Factor Structure of the French Version of the Wechsler Adult Intelligence Scale-III. [Peer Reviewed]. *Educational and Psychological Measurement*. 2004;64(3):463–474.
- Gregorich SE. Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. [Peer Reviewed]. *Medical Care*. Special Issue: Measurement in a multi-ethnic society. 2006;44(11, Suppl 3):S78–S94.
- Haan MN, Mungas DM, Ortiz TA, Acharya A, González HM, Jagust WJ. Prevalence of dementia in older latinos: The influence of type 2 diabetes mellitus, stroke and genetic factors. *Journal of the American Geriatrics Society*. 2003;51:169–177. [PubMed: 12558712]
- Hinton L, Carter K, Reed BR, Beckett L, Lara E, DeCarli C, et al. Recruitment of a community-based cohort for research on diversity and risk of dementia. *Alzheimer's Disease and Associated Disorders*. In Press.
- Insua AM. WAIS-R factor structures in two cultures. [Peer Reviewed]. *Journal of Cross-Cultural Psychology*. 1983;14(4):427–438.
- Jöreskog, KG.; Sörbom, D. LISREL 8: User's reference guide. Chicago: Scientific Software International; 1996.
- Kaufman AS, McLean JE, Reynolds CR. Analysis of WAIS ÄiR factor patterns by sex and race. [Peer Reviewed]. *Journal of Clinical Psychology*. 1991;47(4):548–557. [PubMed: 1939700]
- Laicardi C, Frustaci A, Lauriola M. Factor structure of the WAIS AiR Italian version compared with American and British solutions. [Peer Reviewed]. *Psychological Reports*. 1996;79(3 Pt 2):1171–1177.

- Larrabee GJ, Kane RL, Schuck JR. Factor analysis of the WAIS and Wechsler Memory Scale: An analysis of the construct validity of the Wechsler Memory Scale. *Journal of Clinical Neuropsychology*. 1983;5(2):159–168. [PubMed: 6863562]
- McArdle JJ. Current directions in structural factor analysis. [Peer Reviewed]. *Current Directions in Psychological Science*. 1996;5(1):11–18.
- Meredith W, Teresi JA. An essay on measurement and factorial invariance. *Med Care*. 2006;44(11 Suppl 3):S69–77. [PubMed: 17060838]
- Mungas D, Reed BR, Crane PK, Haan MN, González H. Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment*. 2004;16(4):347–359. [PubMed: 15584794]
- Mungas D, Reed BR, Haan MN, González H. Spanish and English Neuropsychological Assessment Scales: relationship to demographics, language, cognition, and independent function. *Neuropsychology*. 2005;19(4):466–475. [PubMed: 16060821]
- Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish neuropsychological tests for older persons. *Neuropsychology*. 2000;14:209–223. [PubMed: 10791861]
- Mungas D, Reed BR, Tomaszewski Farias S, DeCarli C. Criterion-Referenced Validity of a Neuropsychological Test Battery: Equivalent Performance in Elderly Hispanics and Non-Hispanic Whites. *Journal of the International Neuropsychological Society*. 2005;11:620–630. [PubMed: 16212690]
- Mungas DM, Reed BR, Jagust WJ, Kramer JH, Weiner MW, Mack WJ, et al. Differential decline in memory and executive functioning in patients with and without subcortical infarcts. [Abstract]. *Neurology*. 2000;54:A73.
- Murray A, Jones DM. Articulatory complexity at item boundaries in serial recall: the case of Welsh and English digit span. *J Exp Psychol Learn Mem Cogn*. 2002;28(3):594–598. [PubMed: 12018511]
- Muthén, LK.; Muthén, BO. *Mplus User's Guide*. 5. Los Angeles, CA: Muthén & Muthén; 1998–2007.
- Olazarán J, Jacobs DM, Stern Y. Comparative study of visual and verbal short-term memory in English and Spanish speakers: testing a linguistic hypothesis. *Journal of the International Neuropsychological Society*. 1996;2(2):105–110. [PubMed: 9375195]
- Orsini A, Laicardi C. Factor structure of the Italian version of the WAIS ÄiR compared with the American standardization. [Peer Reviewed]. *Perceptual and motor skills*. 2000;90(3 Pt2):1091–1100. [PubMed: 10939053]
- Pedraza O, Lucas JA, Smith GE, Willis FB, Graff-Radford NR, Ferman TJ, et al. Mayo's older African American normative studies: confirmatory factor analysis of a core battery. *Journal of the International Neuropsychological Society*. 2005;11(2):184–191. [PubMed: 15962706]
- Ponton MO, Gonzalez JJ, Hernandez I, Herrera L, Higareda I. Factor analysis of the Neuropsychological Screening Battery for Hispanics (NeSBHIS). *Appl Neuropsychol*. 2000;7(1):32–39. [PubMed: 10800626]
- Population Reference Bureau. *Older Americans in the 1990s and Beyond*. 2000
- Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*. 1993;114:552–566. [PubMed: 8272470]
- Smith GE, Ivnik RJ, Malec JF, Kokmen E, Tangalos EG, Kurland LT. Mayo's Older Americans Normative Studies (MOANS): Factor structure of a core battery. *Psychological Assessment*. 1992;4(3):382–390.
- Smith GE, Ivnik RJ, Malec JF, Tangalos EG. Factor structure of the Mayo Older Americans Normative Sample (MOANS) core battery: Replication in a clinical sample. *Psychological Assessment*. 1993;5(1):121–124.
- Steiger JH, Shapiro A, Browne MW. On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*. 1985;50:253–264.
- Stigler JW, Lee S-y, Stevenson HW. Digit memory in Chinese and English: Evidence for a temporally limited store. *Cognition*. 1986;23(1):1–20. [PubMed: 3742988]

- Tuokko HA, Chou PH, Bowden SC, Simard M, Ska B, Crossley M. Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *J Int Neuropsychol Soc.* 2009;15(3):416–425. [PubMed: 19402928]
- Day, JC. U.S. Bureau of the Census. *Population Projections of the United States, by Age, Sex, Race, and Hispanic Origin.* Washington DC: U.S. Government Printing Office; 1993. p. P25-1104.
- Unverzagt FW, Gao S, Baiyewu O, Ogunniyi AO, Gureje O, Perkins A, et al. Prevalence of cognitive impairment: data from the Indianapolis Study of Health and Aging. *Neurology.* 2001;57(9):1655–1662. [PubMed: 11706107]
- Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods.* 2000;3(1):4–69.
- Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain Bryant. In: Kendall, J.; Michael, Windle; West Stephen, G., editors. *The science of prevention: Methodological advances from alcohol and substance abuse research.* Washington, DC, US: American Psychological Association; 1997. p. 1997p. 1281-1324.p. xxxiip. 1458

Table 1

Demographic Characteristics of Study Sample

	African American (n=443)	Caucasian (n=760)	Hispanic English (n=451)	Hispanic Spanish (n=882)	Combined (n=2536)
Gender N (%) Female	305 (68.8)	414 (54.5)	263 (58.3)	553 (62.7)	1535 (60.5)
Education (years) Mean (S.D.) Range	12.9 (2.9) 3–21	14.0 (3.3) 0–21	11.0 (4.2) 0–21	4.7 (4.4) 0–21	10.1 (5.6) 0–21
Age (years) Mean (S.D.) Range	73.8 (7.3) 60–99	74.7 (7.6) 60–96	71.0 (7.1) 60–91	71.6 (7.5) 60–96	72.8 (7.6) 60–99

Table 2

Hypothesized models and constituent factors and indicators

MODEL	COMMON FACTORS
1-factor	global cognition
2-factor	episodic memory vs. non-memory
4-factor	episodic, semantic/language, spatial, attention/working memory/fluency
4-factor	episodic, semantic/language/fluency, spatial, attention/working memory
4-factor	episodic, semantic/language/fluency, spatial, attention/working memory/fluency
5-factor	episodic, semantic/language, spatial, attention/working memory, fluency
6-factor	episodic, semantic, language, spatial, attention/working memory, fluency
7-factor	episodic, semantic, language, spatial, attention/working memory, fluency, abstraction

Table 3

Standardized factor loadings (standard errors in parentheses) from final 5-factor multiple group CFA that freely estimated non-invariant parameters (non-invariant parameters are bolded and underlined).

Factor - SENAS Measure	Caucasian	African American	Hispanic- English	Hispanic- Spanish
Episodic Memory				
Word List Learning - 1	0.89 (0.03)	0.85 (0.03)	0.87 (0.03)	0.85 (0.03)
Word List Learning - 2	<u>0.92</u> (0.03)	0.91 (0.03)	0.92 (0.03)	0.91 (0.03)
Spatial Configuration Learning	0.36 (0.04)	0.30 (0.01)	0.30 (0.03)	0.29 (0.01)
Semantic/Language				
Picture Association	0.76 (0.03)	<u>0.70</u> (0.01)	0.81 (0.03)	0.81 (0.02)
Object Naming	0.74 (0.03)	0.81 (0.02)	0.82 (0.03)	0.85 (0.02)
Verbal Comprehension	0.49 (0.03)	0.60 (0.01)	0.59 (0.04)	0.63 (0.02)
Verbal Expression	0.42 (0.02)	<u>0.52</u> (0.01)	0.54 (0.02)	0.58 (0.01)
Verbal Conceptual Thinking	0.73 (0.03)	0.77 (0.01)	0.78 (0.03)	<u>0.87</u> (0.02)
Spatial				
Pattern Recognition	0.70 (0.03)	0.70 (0.02)	0.76 (0.03)	<u>0.85</u> (0.02)
Spatial Localization	0.72 (0.03)	0.72 (0.02)	0.78 (0.03)	0.73 (0.02)
Non-Verbal Conceptual Thinking	0.69 (0.03)	0.76 (0.02)	0.75 (0.03)	0.76 (0.02)
Spatial Configuration Learning	0.43 (0.03)	0.44 (0.01)	0.50 (0.04)	0.51 (0.01)
Attention/Working Memory				
Working Memory	0.95 (0.03)	0.95 (0.02)	0.96 (0.03)	0.95 (0.02)
Verbal Attention	0.56 (0.03)	0.51 (0.01)	0.61 (0.03)	0.66 (0.01)
Visual Attention	0.57 (0.03)	0.58 (0.01)	0.62 (0.03)	<u>0.70</u> (0.02)
Verbal Comprehension	<u>0.41</u> (0.04)	0.24 (0.01)	0.25 (0.02)	0.22 (0.01)
Fluency				
Phonemic Fluency	0.75 (0.03)	0.69 (0.02)	0.75 (0.03)	0.76 (0.02)
Category Fluency	0.81 (0.03)	0.76 (0.02)	0.81 (0.03)	0.82 (0.03)