Published in final edited form as:
Neuropsychology. 2010 May ; 24(3): 402–411. doi:10.1037/a0017515.

Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers?

Karen L. Siedlecki, **Jennifer J. Manly**, **Adam M. Brickman**, **Nicole Schupf**, **Ming-Xin Tang**, and **Yaakov Stern**
Columbia University Medical Center

Abstract

Objective—The purpose of this study was to examine whether neuropsychological tests translated into Spanish measure the same cognitive constructs as the original English versions.

Method—Older adult participants (N = 2,664), who did not exhibit dementia from the Washington Heights Inwood Columbia Aging Project (WHICAP), a community-based cohort from northern Manhattan, were evaluated with a comprehensive neuropsychological battery. The study cohort includes both English (*n* = 1,800) and Spanish speakers (*n* = 864) evaluated in their language of preference. Invariance analyses were conducted across language groups on a structural equation model comprising four neuropsychological factors (memory, language, visual-spatial ability, and processing speed).

Results—The results of the analyses indicated that the four-factor model exhibited partial measurement invariance, demonstrated by invariant factor structure and factor loadings but nonequivalent observed score intercepts.

Conclusion—The finding of invariant factor structure and factor loadings provides empirical evidence to support the implicit assumption that scores on neuropsychological tests are measuring equivalent psychological traits across these two language groups. At the structural level, the model exhibited invariant factor variances and covariances.

Keywords

aging; cognition; cross-cultural; neuropsychology; measurement invariance; structural equation modeling

A critical consideration in conducting cross-cultural research is the examination of measurement invariance. *Quantitative* differences in cognition, often in the form of mean performance, are routinely reported, but few researchers have examined the *qualitative* differences between groups of interest (but see Edwards & Oakland, 2006; Dolan, 2008). Invariance analyses are tools that allow researchers to examine whether the variables of interest represent the same theoretical constructs across groups. The establishment of measurement invariance provides evidence for the assumption that scores on tests measure

K.L. Siedlecki, Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center; J.J. Manly and A.M. Brickman, Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Gertrude H. Sergievsky Center, and Department of Neurology, Columbia University Medical Center; N. Schupf, Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Gertrude H. Sergievsky Center and Department of Epidemiology, School of Public Health, Columbia University Medical Center and NYS Institute for Basic Research, Staten Island, NY; M-X. Tang, Gertrude H. Sergievsky Center and Department of Biostatistics, School of Public Health, Columbia University Medical Center; Y. Stern, Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Gertrude H. Sergievsky Center, and Departments of Neurology and Psychiatry, Columbia University Medical Center.

equivalent psychological traits among diverse groups. Establishing the equivalence of neuropsychological constructs used for diagnosing dementia across different cultural and linguistic groups has become increasingly important as the percentage of adults who are at risk for developing dementia increases and a greater proportion of these individuals come from diverse cultural backgrounds.

Neuropsychological batteries used in the cognitive assessment of individuals being evaluated for disorders associated with increased age, such as Alzheimer's disease, are typically developed and validated in English speaking individuals. Although translated batteries are administered to non-English speakers under the common assumption that the battery has the same meaning across language groups, research has indicated that test scores often result in the over-diagnosis of cognitive disorders in non-English speakers. This problem is particularly true among elderly Spanish speaking Hispanics, who are a growing demographic group in the United States. Even on basic tests of cognition, such as a translated version of the Mini-Mental State Exam (MMSE), Spanish-speaking subjects are more likely to be categorized as impaired, despite a normal clinical evaluation (Bird, Canino, Stipec, & Shrout, 1987).

The measurement invariance of cognitive abilities has been extensively examined across different age groups both cross-sectionally (e.g. Bowden, Weiss, Holdnack, & Lloyd, 2006; Schaie, Willis, Jay, & Chipuer, 1989; Taub, McGrew, & Witta, 2004) and longitudinally (e.g. Hertzog & Schaie, 1986). Researchers have also evaluated the measurement invariance of cognitive abilities across sex (e.g. Maitland, Intrieri, Schaie, & Willis, 2000) and across groups with differing clinical presentations (e.g. Siedlecki, Honig & Stern, 2008). Although a few researchers have examined measurement invariance of cognitive abilities across race (Edwards & Oakland, 2006; Dolan, 2000) only one, to our knowledge, has examined measurement invariance of cognition across language groups (Tuokko, Chou, Bowden, Simard, Ska, & Crossley, 2009). Tuokko et al. (2009) found evidence of partial measurement invariance across English and French speakers on a neuropsychological battery from the Canadian Study of Health and Aging.

One method used to evaluate measurement invariance across groups is multi-group confirmatory factor analysis (CFA). Using structural equation modeling (SEM) we can determine whether the corresponding relationships among the variables and constructs are the same across English and Spanish speakers in a community-based cross-cultural sample. There are several types of invariance analyses, often accompanied by different terminology (Vandenberg & Lance, 2000). In this study we adopted the guidelines proposed by Widaman and Reise (1997), but the terminology recommended by Vandenberg and Lance (2000), and examined three increasingly stringent levels of measurement invariance- configural invariance (Horn, McArdle, & Mason, 1983), metric invariance (Horn & McArdle, 1992; labeled weak factorial invariance by Widaman & Reise, 1997), and scalar invariance (denoted strong factorial invariance by Meredith, 1993; Widaman & Reise, 1997). A fourth level of measurement invariance, termed "strict" metric invariance (Meredith, 1993) involves constraining the residual variances across groups. However, it is widely accepted that such constraints on error parameters are overly restrictive (Bryne, 2004) and as such, we did not evaluate strict metric invariance. At the structural level, we examined invariant factor variances and covariances.

Configural invariance requires that the pattern of relationships, or the factor structure, is identical across groups. That is, each factor is associated with the same set of items across the groups. Configural invariance is evaluated by examining the fit of the multi-group model. As recommended by Hu & Bentler (1998), multiple fit indexes are examined to evaluate an overall pattern of fit, rather then focusing on just one goodness-of-fit statistic.

Therefore, the chi-square (X^2), the critical ratio (X^2/df), and the root mean square error of approximation (RMSEA), for which numbers closer to zero indicate a better fit, are all reported. RMSEA values < .06 are typically indicative of a good fit (Hu & Bentler, 1999), values between .08 and .10 are generally indicative of a mediocre fit (MacCallum, Browne, & Sugawara, 1996) and values > .10 are usually considered to be indicative of a poor fit. Bentler's comparative fit index (CFI) and the Tucker Lewis Index (TLI) were also examined and for these fit statistics, values closer to 1.0 indicate a better fit (Hu & Bentler, 1999). Specifically, values > .95 are considered to signify a good fit (Hu & Bentler, 1999), although sometimes a cut-off of > .90 is used (Bentler, 1992).

Metric invariance is established if the magnitudes of the unstandardized factor loadings are the same across the groups. Metric invariance provides evidence that the corresponding latent factors have the same meaning across the groups of interest, because the latent factors reflect the common, or shared, variance among the observed variables. If the corresponding latent variables reflect the same meaning across the groups we would expect the relationship between the observed variables and the latent variables to be the same.

Scalar invariance is evaluated by constraining the intercepts of the measured variables to be the same across groups.

Separate from measurement invariance is structural invariance. *Structural invariance* refers to the invariance of the relationships between or among the latent variables. It is worthwhile to test for invariance of structural parameters only if measurement invariance (or partial measurement invariance) is obtained. In this study, we examined whether the variances and covariances among the corresponding latent constructs are equivalent in magnitude.

Each subtype of invariance is tested by constraining the corresponding parameters to be equal across the groups, and comparing the fit of the constrained model to the previous model. If the constrained model does not fit substantially worse than the previous invariance model then it may be argued that the model demonstrates invariance (i.e., the hypothesis of invariance cannot be rejected). The change in chi-square per change in degrees of freedom between the models is typically used to determine whether the fit of models are significantly different. However, it is well known that the chi-square statistic is affected by sample size such that large differences with a small sample may not be significant, but small or trivial differences in a model with a large sample size may yield a highly significant chi-square. As such, the change in several fit statistics (CFI, TLI, and RMSEA) is also evaluated to supplement the $\Delta X^2/\Delta\, df$ test. Recent work by Cheung and Rensvold (2002) suggests that the CFI statistic, in particular, may be valuable in determining changes in fit. It is recommended that if the change in the CFI equal to or less than − .01, than the invariance hypothesis should not be rejected, a change in the CFI value greater than −.01 would indicate the differences between the model fits are substantial (Vandenberg & Lance, 2000).

The purpose of this study was to examine whether the factor structure of a set of neuropsychological tests exhibited invariance at the measurement and structural level across English and Spanish speakers. The establishment of measurement invariance provides evidence for the assumption that scores on the neuropsychological tests measure equivalent psychological constructs among the groups, and renders the interpretation of quantitative comparisons uncomplicated. If metric invariance is found lacking then this would create issues in the interpretation of quantitative differences in individual variables.

Method

Participants

The sample included participants from the Washington Heights and Inwood Columbia Aging Project (WHICAP), a prospective, community-based epidemiological study of aging and dementia in northern Manhattan, the details of which have been described in earlier work (Manly, Bell-McGinty, Tang, et al., 2005; Tang, Cross, Andrews, et al., 2001). The study combined data from two recruitment efforts in this community - one of which began in 1992 and the other in 1999. In brief, a stratified random sample of 50% of all individuals older than 65 years was obtained through the Center for Medicare Services (CMS). The CMS sent letters informing the individuals that they had been selected to be a participant in a study of aging through Columbia University. Individuals who responded to the letter, a phone call to their home, or a visit to their address, and were willing to be included as participants, underwent an in-person interview assessing general health, functional ability and their medical history, in addition to a neuropsychological battery. A physical and neurological examination was also conducted. Ethnicity was classified by self-report (vis-à-vis the 1990 US Census guidelines) in which the participant was first asked whether they were Hispanic and Latino, and then in a separate question, asked to classify themselves as white, black, Asian, American Indian/Pacific Islander, or other. Evaluations were conducted in English or Spanish, based on the participant's language of preference. This research was approved by local institutional review boards.

Only data from the baseline visit of cognitively-healthy adults were included in the current study. Those individuals diagnosed with dementia (n = 52 in English speakers; n = 31 in Spanish speakers) or "questionable dementia" at baseline were excluded (n = 371 in English speakers; n = 360 in Spanish speakers). Questionable dementia was determined by clinical consensus when a patient had either sufficient cognitive impairment for a diagnosis of dementia but no functional impairment or had insufficient cognitive impairment for a dementia diagnosis but had been assigned a Clinical Dementia Rating of 0.5 by the examining neurologist because of some functional impairment. In addition, the data from participants with major medical, neurological (e.g., stroke, depression, brain tumor, epilepsy, Parkinson's disease, Korsakoff's syndrome), or significant psychiatric disorders (e.g., depression) were excluded from the analyses (n = 85 in English speakers; n = 61 in Spanish speakers). The final sample (N = 2664) was composed of 1800 English speakers and 864 Spanish speakers, demographic characteristics of which are presented in Table 1.

In the neighborhoods of Washington Heights and Inwood many businesses and institutions are bilingual or exclusively Spanish-speaking. As a result many of the older Hispanic adults in the cohort speak little to no English. The English speakers were significantly older and had more years of education than the Spanish speakers in the sample. The English speakers predominately spoke English at home, whereas the Spanish speakers predominately spoke Spanish at home. Those individuals whose neuropsychological evaluation was conducted in Spanish, were classified as Spanish-speakers.

Neuropsychological Assessment

All participants were administered a brief neuropsychological battery designed to assess a broad range of cognitive functioning such as memory, language, visual-spatial ability, and reasoning. The battery comprises subtests from widely-used standardized neuropsychological tests (see Stern et al., 1992, for details on the development of the core battery). Participant performance on these tasks is presented in Table 2. The assessment battery is conducted in English and Spanish by balanced bilingual research staff who speaks both daily with friends, family, and colleagues. At the first assessment, participants are

asked their opinion of which language would yield their best performance, and this language is used to administer the battery throughout all follow-up visits. All interview questions, test instructions, and stimuli were translated into Spanish by a committee of Spanish speakers from Cuba, Puerto Rico, Spain, and the Dominican Republic, and then back-translated to ensure accuracy. Where necessary, scoring criteria were modified so that credit is given for responses reflecting regional idioms (Jacobs et al, 1997).

In the Selective Reminding Test (SRT; Buschke & Fuld, 1974), participants are read a list of 12 words six times and after each of the six trials they are asked to recall the words. After each recall attempt, participants are reminded of the words they failed to recall. *SRT total recall* refers to the total number of words out of a possible 72 (12 words x 6 trials) that the participant correctly recalled. After a 15-min delay, participants are asked to recall the 12 words. The *SRT delayed recall* score refers to the number correct (out of 12). After the delayed recall portion, participants are administered an SRT delayed recognition test in which each of the 12 words are presented with three distracters. *SRT delayed recognition* is number of words correctly recognized.

In the modified 15-item Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983), participants are presented with 15 line drawings and asked to identify each object. If a participant is unable to name the object, the examiner gives the participant a semantic cue after 20 s, followed by a phonemic cue after 15 s. The *Naming total* variable refers to the total number of objects named spontaneously.

Two tests of verbal fluency were administered. In the *Letter Fluency* test, participants are given three letters (i.e., *C*, *F*, *L*) and asked to generate as many words as they can that begin with each letter in 60 s (within specific guidelines). Total number of words named across the three letters was used as the score. In the *Category Fluency* test, participants are given a category (e.g., animals) and asked to generate as many items as they can that are a member of the given category in 60 s. The total number of words generated across the categories was used as the score.

The Benton Visual Retention Test (BVRT; Benton, 1955) is comprised of two parts- the recognition test and the matching test. In the *BVRT recognition* test, participants view a nonverbal design for 10 s and are then asked to select the design from an array with three distracters. In the *BVRT matching* test, participants are asked to match each nonverbal design to an identical design in an array of four smaller designs. In both cases, the total number correct was used as the score.

In the *Rosen* Drawing Test (Rosen, 1981), participants copy five visual designs onto a piece of paper. No partial credit is given, and drawings are scored as either correct or incorrect. The *Rosen* variable refers to the total number of designs correctly copied.

The *Similarities* test is a subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) and requires participants to articulate similarities in a set of items. The total raw score was used in the analyses.

The *Identities and Oddities* test is a subtest of the Mattis Dementia Rating Scale (Mattis, 1976) and requires the participant to examine three items and select which two are alike in the first eight trials. In the second eight trials, the same items are shown and the participant is required to select which item is different. The total number of items correct across trials was used as the measure.

The *Repetition* task is a subtest of the Boston Diagnostic Aphasia Evaluation (BDAE; Goodglass, 1983), which requires participants to repeat phrases read by the examiner. Only

the high probability phrases were used. The total number of phrases correct was used as the score.

The *Comprehension* test of the BDAE requires participants to answer basic comprehension questions.

Only a subset of the sample completed the *Color trials 1* (n = 955) and *Color trails 2* (n = 919) tests designed to assess processing speed. The Color Trails test (*CTT*) requires participants to connect numbers (*CTT 1*) or numbers alternating in the same color (*CTT 2*) in the appropriate order as quickly as possible. Note that in these two tasks higher scores indicate slower speeds.

Analyses

All confirmatory and invariance analyses were conducted with Amos 16.0 (Arbuckle, 2003) using raw scores. Full-information maximum likelihood estimation was used to deal with missing data.

To identify the underlying factor structure, an exploratory factor analysis[1] (EFA) using principal axis factoring and oblique rotation was performed on the 15 variables of interest in the English speaking sample only. The model derived from the EFA was converted to a simple structure CFA model, in which each variable loaded only on the factor that it had the highest loading.

Invariance analyses were conducted to assess whether the structural model was measuring the same constructs across the English and Spanish speakers. Configural invariance, in which the structure of the model was constrained to be the same across the two language groups, was assessed by examining the overall fit of the model. Metric invariance, in which the magnitude of the corresponding factor loadings were constrained to be equal across the two language groups, was assessed by comparing the metric invariance model fit to the configural invariance model fit. Scalar invariance was evaluated by constraining corresponding intercepts to be the same across groups. Two additional models, examining aspects of structural invariance, included additional constraints on factor variances and covariances.

Results

The means and standard deviations of each task are presented in Table 2 separately for the English and Spanish speakers. English speakers obtained higher scores on each of the 15 tasks than did the Spanish speakers. On average, the English speakers had more than 5 additional years of formal education than did the Spanish speakers (see Table 1). Inspection of the correlation matrix presented in Table 3 shows that nearly all the variables were correlated with one another, and education was significantly associated with all the variables.

[1]Oftentimes researchers use a principal components analysis (PCA) when attempting to identify the underlying factor structure of a set of variables. Although PCA and EFA are very similar in that they are both performed by examining the pattern of correlations between the observed measures, PCA does not differentiate between the common and unique variance among the observed measures (whereas EFA does). As such, EFA is better equipped to identify factors that represent what is common among the variables and PCA is more appropriately used as a data reduction technique (see Fabrigar, Wegener, MacCallum, & Strahan, 1999, for a comprehensive review of EFA in psychological research).

Factor analysis

Typically, the first step in invariance analyses is to identify a baseline model that provides a good fit to the data for each group (Byrne et al., 1998; Bowden et al., 2008). The English-speaking sample was used in the EFA to generate a factor structure that could be used as a baseline against which to compare the Spanish-speaking sample to. Only English-speaking participants without missing data (n = 899) were used (a second EFA was performed with all English speakers, replacing missing data with the mean value, and the subsequent factor structure had no substantive differences).

Several methods were used to determine the number of factors to retain. The scree plot was inspected but a disproportionate amount of the variance was accounted for by the first factor. Adherence to the Kaiser eigenvalue > 1 rule resulted in four factors. Although solutions obtained by retaining the number of factors indicated by the eigenvalue > 1 rule can often lead to over-factoring, inspection of the factor solution showed that the four retained factors were interpretable and consistent with prior research on the neuropsychological variables in a different sample (see Siedlecki et al., 2008).

From the EFA, the factors of memory, language, processing speed, and visual-spatial ability were identified. The three SRT variables loaded on the memory factor. The language factor was comprised of the naming total variable, the category and letter fluency tests, the WAIS-R similarities subtest, and the BDAE repetition and comprehension subtests. The processing speed factor comprised two timed tasks, color trails 1 and color trails 2. The BVRT recognition and matching variables, the Rosen Drawing Test, and the Identities and Oddities subtest loaded on the visual-spatial ability factor.

This four-factor model derived from the EFA was converted to a CFA model (see Figure 1). This four-factor model fit the data well in both the English and Spanish samples, and across the total sample (see Table 4). To ensure that the four-factor model was the most appropriate representation of the data, the fit of a three-factor and five-factor model (the structures derived from an EFA specifying x factors) was examined across the English speakers, Spanish speakers, and the total sample. Inspection of Table 4 indicates that the four-factor model was the best-fitting model within each sample, both in relative and absolute terms.

The standardized coefficients and correlations in the four factor model for the English speakers, Spanish speakers, and total sample are presented in Table 5. All the path coefficients were significantly different from zero at the $p < .001$ level, providing evidence of convergent validity. The inter-factor correlations were also all significant, but even the largest correlations were substantially less than 1.0, thereby providing evidence of discriminant validity.

Invariance Analyses

Once it was established that a four-factor model comprising memory, language, speed and visual-spatial ability constructs fit the data well for both language groups, invariance analyses were conducted across the English-speaking and Spanish-speaking groups. Configural invariance was first evaluated by specifying the structure of the model (see Figure 1) to be the same across the two groups. The fit of the configural model was good (X^2= 1022.27, df = 168; RMSEA = .04; CFI= .93), suggesting that the four-factor model was an appropriate representation of the data across both groups.

Metric invariance was examined by constraining the corresponding factor loadings to be equivalent across the English and Spanish speakers and comparing the fit of the metric invariance model to the configural invariance model. As can be seen in Table 6 (Model 2), the change in chi-square per change in df was significant, indicating that the metric

invariance model fit significantly worse than the configural model. However, inspection of the change in CFI indicates that the change was not substantial (ΔCFI = .014) and the change falls within the guidelines proposed by Cheung and Rensvold (2002) for not rejecting the invariance hypothesis. Further, the overall fit of the metric invariance model was quite good (CFI = .917; RMSEA = .046). In the next step, scalar invariance was examined by additionally constraining the observed score intercepts to equality across the groups. The fit of this model (Model 3a, Table 6) was significantly worse than the preceding model (Model 2, Table 6), and the CFI, TLI, and RMSEA all showed appreciable reductions in fit. Evaluation the fit indices for each variable intercept when constrained separately, indicated that the largest reductions in fit were associated with the constraints on the variables loading on the Language factor (i.e., naming total, letter fluency mean, category fluency mean, comprehension, and similarities), the Speed factor (i.e., CTT1 and CTT2) and the Visual-Spatial factor (i.e., BVRT recognition and matching), as well as the SRT total recall variable. When the intercepts of these ten variables were unconstrained, there was no appreciable change in fit of the model (Model 3b, Table 5). Therefore partial measurement invariance was obtained by allowing these ten intercepts to vary across the groups.

Aspects of structural invariance (the examination of the relations among the latent variables) were examined by placing additional constraints on Model 3b. In Model 4 (Table 6) corresponding latent variances were constrained across the groups. Although the $\Delta X^2/\Delta\, df$ was significant, there was no appreciable change in fit of the other indices. In Model 5 corresponding covariances were constrained to equality across groups yielding a non-significant change in chi-square.

It is unsurprising that we found evidence of only partial scalar invariance. Scalar invariance is demonstrated by equivalent manifest variable intercepts. However, it is clear that there are group differences in the mean of observed variables across English and Spanish speakers (see Table 2). These differences can be partially accounted for by the significantly lower educational attainment of the Spanish speakers (M= 7.01, SD = 4.24) as compared to the English speakers (M= 12.48, SD= 3.75) in our sample. In cases such as this, Vandenberg and Lance (2000) argue, “a test for intercept or scalar invariance (i.e., no differences between groups) is not appropriate because difference in item location parameters would be fully expected. However, these differences are not biases in the sense of being undesirable as in rating source biases, but rather they reflect expected group differences.” (pg 38).

To further investigate the intercept differences in the sample, post-hoc invariance analyses were conducted with the previously described four-factor model across education-matched subsamples of English and Spanish speakers (Bowden et al., 2008). In these analyses only English speakers (n = 539) with educational attainment of 11 years or less and Spanish speakers with greater than 3 years of education (n = 664) were included. The mean educational attainment (M = 8.12, SD = 2.47) for this English speaking subsample was not significantly different than the Spanish subsample (M= 8.56, SD = 3.55). Because of the number of comparisons a p level of < .01 was used in independent sample t-tests examining the differences in the neuropsychological tests across the English and Spanish speaking subsamples. The Spanish-speaking subsample obtained significantly higher scores on the SRT delayed recall test. The English-speaking subsample obtained better scores on the tests of comprehension, BVRT recognition, BVRT matching, CTT1 and CTT2 (see Table 7).

Results of the invariance analyses are presented in Table 8. The configural invariance model fit the data well, and constraining the factor loadings to equality across groups resulted in a significant $\Delta X^2/\Delta df$ but no appreciable decrements in other fit indices (Model 2, Table 8). Constraining all the intercepts to be invariant across the groups resulted in a significantly worse fit ($\Delta\, X^2$= 201.57, $\Delta\, df$ = 15, p < .01) and the change in CFI, TLI, and RMSEA were

all substantial. Partial scalar invariance was established by freeing the equality constraints on the intercepts for the comprehension subtest, and the BVRT recognition and matching subtests on the Visual-spatial factor. Whereas ten variable intercepts needed to be unconstrained across the groups in the previous invariance analyses, in these analyses only three variable intercepts needed to be unconstrained to obtain partial invariance. Aspects of structural invariance were assessed by constraining the latent variable variances and covariances (Model 4 and Model 5, Table 8) to equality across the groups. The $\Delta X^2/\Delta df$ was significant for Model 4, but not Model 5, and for both models the other fit indices showed no appreciable change, indicating the latent variable variances and covariances were invariant.

Discussion

This study takes an important step in validating the use of measures that were originally developed in English and then adapted for use among non-English speaking samples. It is one of the few studies to examine explicitly the measurement invariance of neuropsychological constructs across language groups (see Tuokko et al., 2009). The results indicate that the neuropsychological tests are likely measuring the same construct across the English and Spanish speaking older adults in this community-based study.

Inspection of the goodness-of-fit statistics for the four-factor model indicated that a model comprising memory, language, visual-spatial ability, and processing speed constructs was a good representation of the data in both language groups, and also across the total sample. This model maps closely to the Cattell-Horn-Carroll (CHC; Carroll, 1993; Flanagan & Harrison, 2005) taxonomy of cognitive abilities. For example, the memory construct is consistent with Glr (i.e., long-term storage and retrieval), the speed construct is analogous with Gs (i.e., processing speed), the visual-spatial ability construct is consistent with Gv (i.e., visual processing) and the language construct relates closely to Gc (i.e, comprehension and knowledge) and perhaps also to Glr which is associated with word fluency and naming facility (McGrew, 2009). Our results are also consistent with our previous study (Siedlecki et al., 2008), which examined the relations among an almost identical set of variables in a sample of cognitively-healthy English speakers recruited from a memory disorders clinic. In Siedlecki et al. (2008) we conducted a series of confirmatory factor analyses to compare alternate models and found that an *a priori* model consisting of memory, language, visual-spatial ability, and speed constructs fit the data the best, along with an attention construct comprised of variables not measured in this study.

Collectively, these findings suggest that the four-factor model depicted in Figure 1 is a good representation of the relations among the neuropsychological variables. The results of the invariance analyses indicate that the structure of the four-factor model and the magnitude of the relationship between the observed variables and the latent constructs are invariant across English and Spanish speakers. The finding of invariance of factor loadings provides empirical evidence to support the assumption that scores on tests measure equivalent psychological traits across these diverse groups. Not surprisingly, the observed score intercepts were not invariant across the English and Spanish speakers. The English speakers obtained significantly higher scores on all the neuropsychological tests as compared to the Spanish speakers. These mean differences likely reflect the significant difference in educational attainment (as well as potential differences in the quality of education). Partial scalar invariance was obtained by allowing 10 of the 15 variable intercepts to vary across the groups. Post-hoc invariance analyses examining language group differences across education-matched subsamples of English speakers and Spanish speakers indicated that there were still intercept differences related to the language and visual-spatial ability factors. Specifically, the comprehension variable from the language factor, and the BVRT

recognition and matching subtests from the visual-spatial ability factor were not invariant across English and Spanish speakers. This finding suggests that these specific variables may be culturally or linguistically dependent.

Our findings of partial measurement invariance is consistent with results reported by Tuokko et al. (2009) who found partial measurement equivalence of French and English versions on a neuropsychological battery. Specifically, they found two of the factors (Long-term Retrieval and Visuospatial speed) displayed invariance and a Verbal Ability factor demonstrated partial scalar invariance, with some observed score intercepts of variables loading on the Verbal Ability factor lacking invariance.

Lack of scalar invariance suggests that factor means can not be meaningfully compared (Chen, 2008; Widaman & Reise, 1997). This would therefore suggest that specific comparisons of the factor means across the English and Spanish-speaking samples are ill-advised. However, evaluation of factor means across education-matched subsamples would less biased, particularly if the non-invariant tasks (i.e., Comprehension, BVRT recognition, BVRT matching) were removed.

In conclusion, each of the latent constructs exhibited configural and metric invariance across the two language groups. Across the full English-speaking sample and the Spanish-speaking sample there was evidence of partial scalar invariance for each of the four constructs. Across the English-speaking sample of individuals with 11 years or less of education and the Spanish-speaking sample with greater than 3 years of education there was evidence of scalar invariance for the memory and speed constructs, and partial scalar invariance for the language and visual-spatial ability constructs. The equivalence of the latent variable variances and covariances indicate invariance of the model at the structural level.

Measurement invariance analyses provide an important framework for investigating how tests and underlying constructs operate across different linguistic and cultural groups (Tuokko et al., 2009) and analyses such as the ones presented here are an important first step in making comparisons across different populations.

Acknowledgments

This work is supported by the National Institute of Aging (P01-AG007232). KLS is supported as a trainee by a grant (T32MH020004-09) from the National Institute of Mental Health.

References

Arbuckle, JL. Amos 5.0. Chicago, IL: SPSS; 2003.

Benton, AL. The Visual Retention Test. New York: The Psychological Corporation; 1955.

Bird HR, Canino G, Stipec MR, Shrout P. Use of the Mini-mental state examination in a probability sample of a Hispanic population. The Journal of Nervous and Mental Disease. 1987; 175:731–737. [PubMed: 3500273]

Bowden SC, Gregg N, Bandalos D, David M, Coleman C, Holdnack JA, Weiss LG. Latent mean and covariance differences with measurement equivalence in college students with developmental difficulties versus the Wechlser Adult Intelligence-III/Wechlser Memory Scale-III Normative Sample. Educational and Psychological Measurement. 2008; 68:621–642.

Bowden SC, Weiss LG, Holdnack JA, Lloyd D. Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale-III. Psychological Assessment. 2006; 18:334–339. [PubMed: 16953736]

Buschke H, Fuld PA. Evaluating storage, retention, and retrieval in disordered memory and learning. Neurology. 1974; 24:1019–1025. [PubMed: 4473151]

Byrne BM. Testing for multigroup invariance using AMOS graphics: A road less traveled. Structural Equation Modeling. 2004; 11:272–300.

Byrne BM, Shavelson RJ, Muthen B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin. 1989; 105:456–466.

Carroll, JB. Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge Press; 1993.

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal. 2002; 9:233–255.

Dolan CV. Investigating Spearman's Hypothesis by Means of Multi-Group Confirmatory Factor Analysis. Multivariate Behavioral research. 2008; 35:21–50.

Edwards OW, Oakland TD. Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. Journal of Psychoeducational Assessment. 2006; 24:358–366.

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods. 1999; 4:272–299.

Flanagan, DP.; Harrison, PL. Contemporary intellectual assessment: Theories, tests, and issues. 2. New York: The Guilford Press; 2005.

Goodglass, H. The assessment of aphasia and related disorders. 2. Philadelphia: Lea & Febiger; 1983.

Hertzog C, Schaie WK. Stability and change in adult intelligence: I. Analysis of longitudinal covariance structures. Psychology and Aging. 1986; 1:159–171. [PubMed: 3267393]

Horn JL, McArdle JJ. A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research. 1992; 18:117–144. [PubMed: 1459160]

Kaplan, E.; Goodglass, H.; Weintraub, S. The Boston Naming Test. Philadelphia: Lea & Febiger; 1983.

Maitland SB, Intrieri RC, Schaie KW, Willis SL. Gender differences and changes in cognitive abilities across the adult life span. Aging, Neuropsychology, and Cognition. 2000; 7:32–53.

Manly JJ, Bell-McGinty S, Tang MX, Schupf N, Stern Y, Mayeux R. Implementing diagnostic criteria and estimating frequency of mild cognitive impairment in an urban community. Archives of Neurology. 2005; 62:1739–1746. [PubMed: 16286549]

Mattis, S., editor. Mental status examination for organic mental syndrome in the elderly patient. New York: Grune & Stratton; 1976.

McGrew KS. CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence. 2009; 37:1–10.

Rosen, W. The Rosen Drawing Test. Bronx, NY: Veterans Administration Medical Center; 1981.

Schaie KW, Willis SL, Jay G, Chipuer H. Structural invariance of cognitive abilities across the adult life span: A cross-sectional study. Developmental Psychology. 1989; 24:652–662.

Siedlecki KL, Honig LS, Stern Y. The structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. Neuropsychology. 2008; 22:400–411. [PubMed: 18444718]

Stern Y, Andrews H, Pittman J, Sano M, Tatemichi T, Lantigua R, Mayeux R. Diagnosis of dementia in a heterogeneous population: Development of a neuropsychological paradigm-based diagnosis of dementia and quantified correction for the effects of education. Archives of Neurology. 1992; 49:453–460. [PubMed: 1580806]

Tang MX, Cross P, Andrews H, Jacobs DM, Small S, Bell K, Merchant C, Lantigua R, Costa R, Stern Y, Mayeux R. Incidence of Alzheimer's disease in African-Americans, Caribbean Hispanics and Caucasians in northern Manhattan. Neurology. 2001; 56:49–56. [PubMed: 11148235]

Taub GE, McGrew KS, Witta EL. A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-Third edition. Psychological Assessment. 2004; 16:85–89. [PubMed: 15023096]

Tuokko HA, Chou PHB, 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. Journal of the International Neuropsychological Society. 2009; 15:416–425. [PubMed: 19402928]

Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods. 2000; 3:4–70.

Wechsler, D. Wechsler Adult Intelligence Scale-Revised. New York: The Psychological Corporation; 1981.

Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: applications in the substance abuse domain. In: Bryant, KJ.; Windle, M., editors. The science of prevention: methodological advance from alcohol and substance abuse research. Washington, DC: American Psychological Association; 1997. p. 281-324.

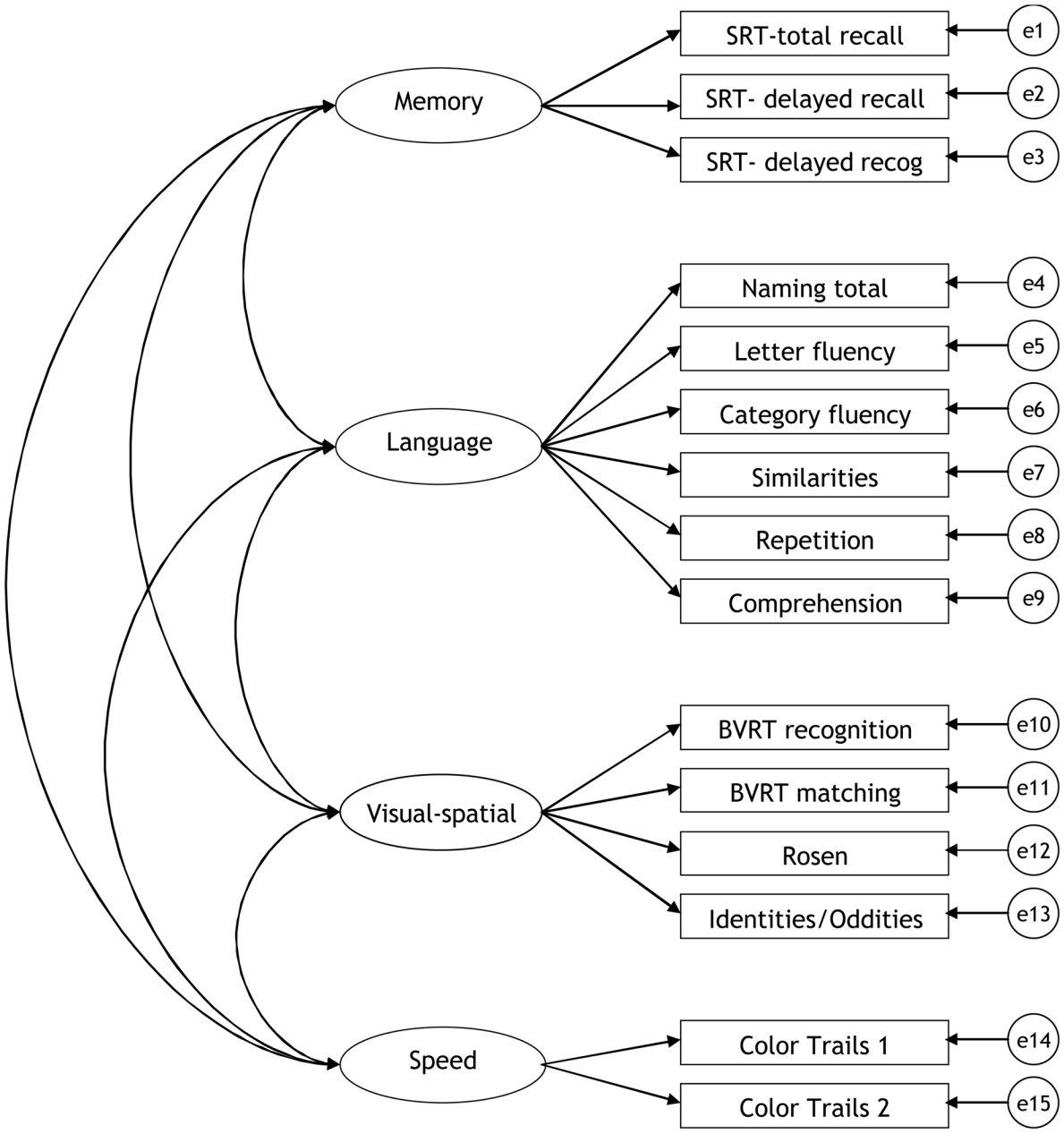

Figure 1.

Representation of the four-factor structural model derived from the EFA and comprised of memory, language, visual-spatial ability and speed constructs. Two-headed arrows connecting latent variables (depicted as circles) represent correlations between the constructs. The paths from the latent constructs to the observed variables (depicted as rectangles) represent the loadings of each task onto its respective construct. The latent variables labeled “e” represent the unique variance and error associated with each observed variable.

Table 1

Sample Characteristics

Demographics			
	English	**Spanish**	**Total**
	Mean (SD)	**Mean (SD)**	**Mean (SD)**
N	1800	864	2664
Age*	75.43 (7.18)	73.61 (6.48)	74.84 (7.01)
Educ(years)*	12.48 (3.75)	7.01 (4.24)	10.70 (4.68)
% Female	66.60	70.10	67.80
Ethnicity			
White, non-hispanic	884 (49.1%)	7 (0.8%)	891 (32.2%)
Black, non-hispanic	786 (43.7%)	6 (0.7%)	792 (29.7%)
Other	33 (5.3%)	1 (0.1%)	34 (1.3%)
Hispanic	96 (5.3%)	850 (98.4%)	946 (35.5%)
Predominant Language Spoken at Home			
English	1385 (76.9%)	32 (3.7%)	1417 (53.2%)
Spanish	26 (1.4%)	571 (66.1%)	597 (22.4%)
Other	73 (4.1%)	3 (0.3%)	76 (2.9%)
Unknown/Missing	316 (17.6%)	258 (29.9%)	574 (21.5%)

Note.

* $p < .01$ between English speaking and Spanish speaking samples

Table 2

Means (and Standard Deviations) on Neuropsychological Tests

	English speakers		Spanish speakers	
Variable	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**
Memory				
SRT- total recall	41.16 (9.57)	0–69	37.62 (7.33)	10–60
SRT- delayed recall	6.17 (2.62)	0–12	5.37 (1.86)	0–12
SRT- delayed recog	11.40 (1.04)	0–12	11.17 (1.15)	4–12
Language				
Naming total	13.76 (1.67)	0–15	13.10 (1.78)	6–15
Letter fluency mean	10.76 (4.47)	0–45	7.56 (3.51)	0 – 27
Category fluency mean	15.73 (4.36)	0–32	12.81 (3.54)	0–31
Similarities	13.73 (6.98)	0–27	8.03 (5.64)	0–25
Repetition	7.71 (.75)	0–8	7.51 (.898)	0–8
Comprehension	5.56 (.86)	0–6	4.51 (1.27)	0–6
Visual-spatial				
BVRT recognition	7.47 (2.08)	0–10	5.77 (2.25)	0–10
BVRT matching	8.89 (1.82)	0–10	7.41 (2.29)	0–10
Rosen	2.82 (1.01)	0–5	2.47 (1.09)	0–5
Identities/Oddities	14.52 (2.43)	0–16	13.48 (2.75)	0–16
Processing Speed				
CTT 1	81.03 (32.29)	23–240	111.77 (47.76)	25–240
CTT 2	145.81 (48.47)	52–240	186.99 (47.61)	69–240

Note. All means are significantly between the English and Spanish speakers at the $p < .001$ level as determined by separate independent-samples *t*-test.

Table 3

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. SRT- total recall	-															
2. SRT- delayed recall	0.72	-														
3. SRT- delayed recog	0.38	0.43	-													
4. Naming total	0.29	0.22	0.20	-												
5. Letter fluency mean	0.44	0.34	0.25	0.39	-											
6. Category fluency mean	0.49	0.38	0.24	0.38	0.64	-										
7. Similarities	0.45	0.35	0.27	0.43	0.62	0.55	-									
8. Repetition	0.20	0.11	0.13	0.27	0.26	0.27	0.25	-								
9. Comprehension	0.26	0.18	0.22	0.36	0.39	0.39	0.46	0.38	-							
10. BVRT recognition	0.32	0.24	0.19	0.33	0.43	0.40	0.49	0.23	0.38	-						
11. BVRT matching	0.25	0.17	0.18	0.36	0.37	0.35	0.42	0.21	0.37	0.66	-					
12. Rosen	0.22	0.16	0.14	0.27	0.31	0.29	0.39	0.16	0.25	0.46	0.46	-				
13. Identities/Oddities	0.17	0.13	0.11	0.24	0.25	0.25	0.31	0.22	0.23	0.50	0.58	0.39				
14. Color trails 1	−0.27	−0.20	**−0.04**	−0.30	−0.41	−0.40	−0.40	−0.20	−0.29	−0.37	−0.38	−0.27	−0.21			
15. Color trails 2	−0.37	−0.29	(−0.09)	−0.30	−0.45	−0.45	−0.47	−0.13	−0.26	−0.39	−0.34	−0.21	−0.26	0.69		
16. Age	−0.29	−0.23	−0.08	−0.11	−0.09	−0.21	−0.14	−0.13	**−.02**	−0.17	−0.13	−0.17	−0.14	0.21	0.19	
17. Education	0.33	0.26	0.23	0.34	0.55	0.43	0.62	0.17	0.42	0.45	0.43	0.33	0.24	−0.36	−0.36	(−0.04)

Note. Bolded values represented correlations that are not significant ($p > .05$), values in parentheses represent correlations that are significant at the $p < .05$ level. All other correlations are significant at the $p < .001$.

Table 4

Goodness-of-Fit Statistics for Structural Models Adapted from Alternative EFA Solutions

	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA	RMSEA 90% CI
English speakers (*n* = 1800)							
3-factor	1900.48	87	21.84	0.80	0.73	0.108	(.103–.112)
4-factor	722.15	84	8.60	0.93	0.90	0.065	(.061–.069)
5-factor	823.11	79	10.42	0.92	0.88	0.072	(.068–.077)
Spanish speakers (*n* = 864)							
3-factor	570.93	87	6.56	0.85	0.79	0.080	(.074–.087)
4-factor	300.13	84	3.57	0.93	0.90	0.055	(.048–.061)
5-factor	332.77	79	4.21	0.92	0.88	0.061	(.054–.068)
Total sample (*n* = 2664)							
3-factor	2274.66	87	26.15	0.85	0.79	0.097	(.094–1.01)
4-factor	846.15	84	10.07	0.95	0.92	0.058	(.055–.062)
5-factor	989.83	79	12.53	0.94	0.90	0.066	(.062–.069)

Table 5

Standardized Path Coefficients and Correlations with 95% Confidence Intervals across Language Group

	Total sample	English speakers	Spanish speakers
N	**2664**	**1800**	**864**
Memory			
SRT- total recall	.90 (.87 – .93)	.92 (.88 – .96)	.80 (.72 – .87)
SRT- delayed recall	.80 (.77 – .84)	.80 (.76 – .85)	.78 (.71– .85)
SRT- delayed recog	.47 (.43 – .51)	.47 (.73 – .52)	.46 (.39 – .53)
Language			
Naming total	.55 (.51 – .59)	.53 (.48 – .58)	.58 (.51 –.65)
Letter fluency mean	.78 (.75 – .81)	.75 (.70 – .79)	.75 (.68 – .81)
Category fluency mean	.75 (.71 – .81)	.77 (.72 – .81)	.57 (.50 – .64)
Similarities	.78 (.75 – .81)	.76 (.72 – .80)	.66 (.58 – .71)
Repetition	.38 (.34 – .41)	.37 (.33 – .42)	.38 (.30 – .45)
Comprehension	.56 (.53 – .60)	.45 (.41 – .50)	.54 (.47 – .60)
Visual-spatial			
BVRT recognition	.81 (.78 – .85)	.79 (.75 – .83)	.73 (.67 – .80)
BVRT matching	.82 (.79 – .86)	.81 (.77 – .85)	.81 (.75 – .87)
Rosen	.59 (.55 – .63)	.54 (.50 – .59)	.65 (.59 – .72)
Identities/Oddities	.65 (.62 – .69)	.70 (.66 – .74)	.56 (.49 – .62)
Processing Speed			
CTT 1	.82 (.77 – .87)	.80 (.73 – .86)	.81 (.69 – .92)
CTT 2	.93 (.88 – .98)	.93 (.87 – .99)	.84 (.73 – .95)
Correlations			
Memory/Language	.62 (.59 – .65)	.65 (.61 – .69)	.46 (.39 – .53)
Memory/Speed	−.47 (−.52 to −.42)	−.49 (−.55 to −.43)	−.23 (−.36 to −.11)
Memory/Visual-spatial	.36 (.32 – .40)	.39 (.34 – .44)	.30 (.22 – .38)
Language/Speed	−.73 (−.77 to −.69)	−.68 (−.73 to −.62)	−.69 (−.78 to −.59)
Language/Visual-spatial	.67 (.64 – .70)	.58 (.54 – .62)	.67 (.62 – .73)
Speed/Visual-spatial	−.71 (−.75 to −.67)	−.57 (−.63 to −.51)	−.73 (−.82 to −.64)

Note. All reported path coefficients and correlations are significantly greater than zero at the $p < .01$ level.

Table 6

Goodness-of-Fit Indices for the Invariance Models for the Four-Factor Model across English speakers (n = 1800) and Spanish speakers (n= 864)

Model	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA (90% CI)	Δ χ^2	Δ *df*	*p* < .01	Δ CFI
Model 1: Configural invariance	1022.27	168	6.08	0.931	0.901	.044 (.041 – .046)				
Model 2: Invariant factor loadings	1208.34	179	6.75	0.917	0.888	.046 (.044 – .049)	186.08	11	yes	−0.014
Model 3a: Model 2 and invariant intercepts	2058.92	194	10.61	0.849	0.813	.060 (.058 – .062)	850.58	15	yes	−0.068
Model 3b: Model 2 and partially invariant intercepts	1377.07	184	7.48	0.903	0.874	.049 (.047 – .052)	168.73	5	yes	−0.014
Model 4: Model 3b and invariant latent variable variances	1513.07	188	8.05	0.893	0.863	.051 (.048 – .053)	135.99	4	yes	−0.010
Model 5: Model 4 and invariant latent variable covariances	1526.27	194	7.87	0.892	0.866	.051 (.048 – .053)	13.20	6	no	−0.001

Table 7

Neuropsychological Performance for Education-matched Subsamples

	English speakers *n* = 539		Spanish speakers *n* = 664	
Variable	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**
Memory				
SRT- total recall	37.15 (8.62)	0–62	38.09 (7.55)	10–60
SRT- delayed recall	5.28 (2.29)	0–12	5.39 (1.92)	0–12
SRT- delayed recog*	11.06 (1.12)	6–12	11.25 (1.10)	4–12
Language				
Naming total	13.16 (1.98)	0–15	13.38 (1.64)	7–15
Letter fluency mean	8.27 (3.64)	0–20	8.29 (3.29)	0 – 27
Category fluency mean	13.58 (3.53)	0–24	13.16 (3.54)	0–31
Similarities	8.79 (6.28)	0–25	8.96 (5.79)	0–25
Repetition	7.56 (.90)	0–8	7.59 (.81)	0–8
Comprehension*	5.31 (1.00)	0–6	4.66 (1.22)	0–6
Visual-spatial				
BVRT recognition*	6.55 (2.27)	0–10	6.10 (2.15)	0–10
BVRT matching*	8.27 (2.12)	0–10	7.82 (2.12)	0–10
Rosen	2.50 (.99)	0–5	2.65 (1.00)	0–5
Identities/Oddities	14.00 (2.66)	0–16	13.63 (2.55)	0–16
Processing Speed				
CTT 1*	92.54 (37.37)	33–240	106.18 (44.09)	31–240
CTT 2*	166.20 (47.58)	80–240	182.34 (47.65)	69–240

Note.

* $p < .01$

Table 8

Goodness-of-Fit Indices for the Invariance Models for the Four-Factor Model across English speakers with educational attainment of 11 years or less (n = 539) and Spanish speakers with educational attainment of greater than 3 years (n = 664)

Model	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA (90% CI)	$\Delta\chi^2$	Δ *df*	*p* < .01	Δ CFI
Model 1: Configural invariance	490.68	168	2.92	0.927	0.895	.040 (.036 – .044)				
Model 2: Invariant factor loadings	516.20	179	2.88	0.924	0.897	.040 (.036 – .044)	25.52	11	yes	−0.003
Model 3a: Model 2 and invariant intercepts	717.77	194	3.70	0.881	0.853	.047 (.044 – .051)	201.57	15	yes	−0.043
Model 3b: Model 2 and partially invariant intercepts	576.18	191	3.02	0.913	0.890	.041 (.037 –.045)	59.98	12	yes	−0.011
Model 4: Model 3b and invariant latent variable variances	590.79	195	3.03	0.910	0.890	.041 (.037 – .044)	14.61	4	yes	−0.003
Model 5: Model 4 and invariant latent variable covariances	600.60	201	2.99	0.909	0.892	.041 (.034 – .044)	9.81	6	yes	−0.001