Title

Testing Measurement Equivalence of Neurocognitive Assessments Across Language in the Hispanic Community Health Study/Study of Latinos

Journal

Neuropsychology, 35(4)

ISSN

0894-4105

Authors

Goodman, Zachary T
Llabre, Maria M
González, Hector M
et al.

Publication Date

2021-05-01

DOI

10.1037/neu0000725

Peer reviewed

HHS Public Access

Author manuscript

Neuropsychology. Author manuscript; available in PMC 2022 May 01.

Published in final edited form as:

Neuropsychology. 2021 May ; 35(4): 423–433. doi:10.1037/neu0000725.

Testing Measurement Equivalence of Neurocognitive Assessments across Language in the Hispanic Community Health Study/Study of Latinos

Zachary T. Goodman, M.A.[1], **Maria M. Llabre, Ph.D.**[1], **Hector M. González, Ph.D.**[2,3], **Melissa Lamar, Ph.D.**[4], **Linda C. Gallo, Ph.D.**[5], **Wassim Tarraf, Ph.D.**[6], **Krista M. Perreira, Ph.D.**[7], **Daniel F. López-Cevallos, Ph.D. M.P.H.**[8], **Priscilla M. Vásquez, Ph.D.**[2], **Luis D. Medina, Ph.D.**[9], **Marisa J. Perera, M.S.**[1], **Donglin Zeng, Ph.D.**[10], **Sierra A. Bainter, Ph.D.**[1]

[1]Department of Psychology, University of Miami, Coral Gables, FL, USA

[2]Department of Neurosciences, University of California San Diego, San Diego, CA, USA

[3]Shiley-Marcos Alzheimer's Disease Research Center

[4]Rush University, Chicago, IL, USA

[5]Department of Psychology, San Diego State University, San Diego, CA, USA

[6]Department of Healthcare Sciences & Institute of Gerontology, Wayne State University, Detroit, MI, USA

[7]Department of Social Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[8]School of Language, Culture, and Society, College of Liberal Arts, Oregon State University, Corvallis, OR, USA

[9]Department of Psychology, University of Houston, Houston, TX, USA

[10]Department of Biostatistics, University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC, USA

Abstract

Objective: Neuropsychological instruments are often developed in English and translated to other languages to facilitate the clinical evaluation of diverse populations or to utilize in research environments. However, psychometric equivalence of these assessments across language must be demonstrated before populations can validly be compared.

Method: To test this equivalence, we applied measurement invariance procedures to a subsample (N = 1,708) of the Hispanic Community Health Survey/Study of Latinos (HCHS/SOL) across English and Spanish versions of a neurocognitive battery. Using cardinality matching, 854 English-speaking and 854 Spanish-speaking subsamples were matched on age, education, sex, immigration status (U.S. born, including territories, or foreign-born), and Hispanic/Latino heritage

Disclosure

Zachary T. Goodman had full access to the study data and take responsibility for the integrity of the data and accuracy of analyses. All authors have reviewed and approved the final manuscript. None of the authors had any financial or other conflicts of interest.

background. Neurocognitive measures included the Six-Item Screener (SIS), Brief-Spanish English Verbal Learning Test (B-SEVLT), Word Fluency (WF) and Digit Symbol Substitution (DSS). Confirmatory factor analysis was utilized to test item-level invariance of the SIS, B-SEVLT, and WF, as well as factor-level invariance of a higher-order neurocognitive functioning latent variable.

Results: One item of both the SIS and WF were more difficult in Spanish than English, as was the DSS test. After accounting for partial invariance, Spanish-speakers performed worse on each of the subtests and the second-order neurocognitive functioning latent variable.

Conclusions: We found some evidence of bias at both item and factor levels, contributing to poorer neurocognitive performance of Spanish test-takers. While these results explain the underperformance of Spanish-speakers to some extent, more work is needed to determine whether such bias is reflective of true cognitive differences or additional variables unaccounted for in this study.

Keywords

Neuropsychological Assessment; Measurement Invariance; Confirmatory Factor Analysis; Cross-Cultural; Language

---

Although Hispanics/Latinos are often viewed as one ethnic minority group, significant heterogeneity exists in sociocultural characteristics among Spanish-speaking populations (Puente & Ardila, 2000). Recent research has suggested differences exist in neuropsychological functioning across Hispanic/Latino background. González and colleagues (2015) reported that compared to 45–74 year old Mexican participants, only South American participants performed equivalently on measures of processing speed and phonemic fluency while Dominican and Central Americans of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) performed worse than their Mexican counterparts. Further, differences across Hispanic/Latino backgrounds were not accounted for when controlling for the effects of age, education, socioeconomic status, or gender on neuropsychological performance (González et al., 2015). In addition, Chilean participants demonstrated poorer performance on phonemic fluency compared to Dominican participants in a separate study, which also balanced groups in terms of age, gender, and years of education (Buré-Reyes et al., 2013). Several possible explanations for these differences in neurocognitive performance exist, including differences in language of assessment and educational attainment, or confounding of heritage group (i.e., country of familial origin) with age, and sex (Buré-Reyes et al., 2013; González et al., 2015). However, research addressing variations in neuropsychological test performance among Hispanic/Latino background groups is a recent shift in the literature, and thus, the explanation for neuropsychological discrepancies remains unclear.

One broadly proposed explanation for differences in neuropsychological performance across ethnic groups is the influence of language (Pedraza & Mungas, 2008). Translational differences or shortcomings can have serious consequences on the psychometric properties and validity of assessments (Artiola i Fortuny & Mullaney, 1997). Accurate translation of assessments from English to another language does not necessarily account for potential

biases, and investigations into other differences, such as item difficulty, are required to establish equivalence in psychometric properties across language (Mungas, Reed, Marshall, & González, 2000). Regional influences on the Spanish dialect of patients from differing Hispanic/Latino heritage groups are potential threats to the validity of neuropsychological measures as well (Artiola i Fortuny & Mullaney, 1997). Moreover, neuropsychological assessments that rely on numerals can introduce linguistic bias on other neurocognitive processes, as differences in the syllables of numbers can increase cognitive load (Hedden et al., 2002). As the proportion of older adults in the United States (U.S.) that is Hispanic/Latino is expected to nearly double and become the second largest older adult population by 2050 (U.S. Census Bureau, 2010), establishing measurement equivalence of tests translated from English to Spanish is especially relevant for modern neuropsychology.

Given the recent awareness of cultural and linguistic factors in neuropsychological assessment, a few studies have directly compared neurocognitive performance across language of assessment. Spanish language of administration has been associated with lower performance on verbal and non-verbal neurocognitive assessments (González et al., 2015) and measures of executive functioning and memory (Brewster et al., 2014). Even tasks not intended to tap verbal ability, such as Trail Making Test (Bezdicek et al., 2016), may be more difficult, as indicated by slower rates of completion, in languages other than English. Surprisingly, Hispanic/Latino English-Spanish bilingual individuals who underwent post-concussion neurocognitive assessment in English performed better than those who were evaluated in Spanish (Ott, Schatz, Solomon, & Ryan, 2014). However, without establishing the measurement equivalence of neuropsychological assessments, it is unknown whether observed language differences reflect true group differences. These discrepancies may result from differences in psychometric properties across language, such as item and test difficulty, as well as systematic differences in level of education, familiarity with standardized testing, or other sociodemographic differences. While one approach to addressing differential performance across the language of assessment is the development of language-specific norms, another method to understand these differences is to identify the psychometric root of differences across language through measurement invariance analyses. Further, psychometric equivalence across language of administration must be established before alternative explanations for such differences can validly be examined.

Measurement invariance refers to psychometric equivalence of an instrument across two or more groups (Meredith, 1993). Typically, invariance is tested using confirmatory factor analysis, by sequentially imposing an increasing number of equality constraints across groups on measurement parameters (Meredith, 1993), resulting in nested models. These constraints force estimated parameters to take the same value across groups, testing the assumption that properties of the test are equivalent. If model fit decreases by a considerable degree, then the assumption of equality is not tenable. Configural invariance tests whether the number of factors, and the items which belong to those factors, are the same across groups. Metric invariance requires factor loadings to be equal across groups. Factor loadings represent how strongly each item contributes to that factor, and violations of metric invariance suggest the factor has different interpretations in each group. Scalar invariance implies that the items are equally difficult, given the same factor score, or latent ability, and is tested by adding equality constraints on item thresholds/intercepts. Thresholds of

dichotomous items are the point along the latent continuum at which a person is more likely to score correct, with higher thresholds indicating more difficult items. In contrast, item intercepts are the score on a continuous item at an average ability level, and lower intercepts generally represent harder items.

In the case of non-invariance at any of the aforementioned stages, partial invariance can be tested by allowing some parameters to vary across groups (Byrne, Shavelson, & Muthen, 1989). To establish a partially invariant model, parameters which significantly decrease model fit when constrained are allowed to vary, while those which do not harm model fit remain constrained. The presence of partial or total non-invariance can be indicative of test bias (Meredith & Teresi, 2006), and these forms of psychometric non-equivalence in clinical neuropsychology can lead to underestimating neurocognitive abilities (Helms, 1992) and incorrectly classifying people diagnostically if left uncorrected (Pedraza & Mungas, 2008). Measurement invariance offers an approach to identifying and understanding language differences that arise as artifacts of measurement rather than true differences in neurocognitive functioning between Spanish and English examinees.

Several studies have examined language-based differences in neuropsychological assessment performance through the application of invariance testing procedures (Mungas, Reed, Crane, Hann, & Gonzáles, 2004; Mungas, Reed, Haan, & González, 2005; Mungas, Widaman, Reed, & Tomaszewski Farias, 2011; Siedlecki et al., 2010; Tuokko et al., 2009). One study demonstrated scalar non-invariance in measures of verbal ability, indicating that in persons of equal verbal comprehension ability, those who underwent assessment in English performed worse than those who were assessed in French on the Similarities subtest of the WAIS-R and the Token Test (Tuokko et al., 2009). The Spanish and English Neuropsychological Assessment Scales (SENAS) was developed simultaneously in Spanish and English to facilitate psychometric equivalence across ethnic groups (Mungas Reed, Marshall, & González, 2000). Mungas and colleagues (2011) reported scalar non-invariance in some verbal and spatial assessments of the SENAS, where some items were more difficult for Hispanics/Latinos who were assessed in Spanish compared to non-Hispanic White and African American participants of the same ability. Similar differences in the difficulty of verbal and spatial assessments between Spanish-speaking and English-speaking participants of otherwise similar ability have been observed (Mungas et al., 2005; Mungas et al., 2004; Siedlecki et al., 2010). Taken together, differences in neuropsychological performance emerge in tests developed across languages (Mungas et al., 2011) as well as tests translated from English (Siedlecki et al., 2010; Tuokko et al., 2009).

While several of the studies discussed above have identified differences across language in neuropsychological performance at the level of a total score, few studies have focused on item-level discrepancies across Spanish and English administration. Whereas invariance at the level of assessments can detect if one language outperforms another, item-level invariance is necessary to detect psychometric discrepancies that lead to biases in assessments. In one study, over half of the items on the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), a widely used general cognitive screener, performed differentially across Spanish and English administration (Jones, 2006), suggesting scalar non-invariance at the item level.

The present study aims to expand on this limited body of research investigating language differences and non-invariance in neuropsychological assessments, specifically at the item level. In this study we tested measurement invariance of a brief neurocognitive battery across Spanish and English administrations within the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Using hierarchical multiple-group confirmatory factor analysis, we examined invariance at both the factor and item levels. To account for other possible explanations of previously observed differences across groups (i.e., age, sex, education, Hispanic/Latino background, & immigration status [U.S.-born vs. foreign-born]), we statistically matched across English and Spanish administrations with respect to these covariates. While additional sociocultural factors, including bilingualism and education quality, are not directly tested in this study, the importance of such considerations in establishing measurement equivalence is further explored in the discussion. The linguistic validation process of this neurocognitive battery is a necessary step to contextualizing previously reported differences across Hispanic/Latino heritage groups.

Method

Participants

This study drew on a sample from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). The HCHS/SOL was designed to investigate risk and protective factors of health, disease, and mortality in the Hispanic/Latino population residing in four cities in the U.S. (Chicago, IL, Miami, FL, the Bronx, NY, and San Diego, CA). The HCHS/SOL includes a wide range of biological, psychosocial, and demographic variables. LaVange and colleagues (2010) as well as Sorlie and colleagues (2010) provide more comprehensive descriptions of the HCHS/SOL study design and sampling method. Briefly, the HCHS/SOL is a multicenter, community-based study of 16,415 self-identified Hispanic/Latino adults. Probability sampling was used to derive a sample representative of the Hispanic/Latino population at each site. A subset of 9,600 participants completed a neurocognitive assessment based on their age at the baseline examination of the HCHS/SOL (2008–2011; Sorlie et al., 2010), which was sampled for the present study.

Neuropsychological testing was conducted by trained, bilingual administrators and participants chose the language in which they were assessed to ensure their comfort throughout the baseline visit. In order to compare participants who were assessed in English versus Spanish on neurocognitive performance, the two groups were matched through cardinality matching (Zubizarreta, Paredes, & Rosenbaum, 2014), with the *designmatch* package in *R*. Briefly, this procedure matches samples based on a set of continuous and categorical covariates to control for potentially confounding differences across groups. Because age and years of education are continuous, a small tolerance of differences in means of each covariate was allowed across groups. In this case, mean differences of .08 standard deviations in age and years of education across language groups was the smallest difference possible to successfully match participants across language, and groups were not significantly different based on independent samples *t*-tests (see Table 1). Spanish and English groups were equally matched on sex, immigration status (U.S. born, including territories, or foreign-born), and Hispanic/Latino heritage group. Missingness on

neurocognitive measures in the full sample ranged from 1.1% to 2.5%. On demographic variables, missingness ranged from 0% (age/sex) to 3.6% (Hispanic/Latino background). Only participants with complete demographic data were included in the analyses to simplify the matching procedure, which resulted in no missingness of neurocognitive variables.

The analytic sample included 1,708 Hispanic/Latino middle-aged adults (60.3% female), ranging from 45 to 74 years old (M = 54.06, SD = 6.77). The sample varied across Hispanic/Latino background (65.2% Puerto Rican, 23.7% Mexican, 4.6% Cuban, 2.5% Dominican, 2.0% Central American, & 2.0% South American). Demographics by language group and statistics of group comparisons are reported in Table 1.

Materials

Six-Item Screener.—The Six-Item Screener (SIS; Callahan, Unverzagt, Hui, Perkins, & Hendrie, 2002) is a brief assessment of global neurocognitive functioning designed to screen for neurocognitive impairment. The SIS includes three short verbal recall items and three orientation questions (i.e., orientation to year, month, and day of the week). Participants are first presented with the three recall items at a rate of one-and-a-half seconds per word. Participants are asked to repeat the words after administering all three items; if the participant incorrectly repeats a word, the words can be repeated two additional times. The test is discontinued if participants fail to correctly repeat the words after three attempts. After repeating all three words correctly, participants are presented the orientation items, and then asked to recall the three previously presented words. Each of the six items is scored as either correct (1) or incorrect (0). Previous psychometric work has demonstrated cut-off scores of 3 provide the highest sensitivity for dementia (Callahan et al., 2002; Xue et al., 2017), performing similarly to lengthier assessments for neurocognitive decline such as the MMSE. Additionally, the reliability of the SIS was adequate (α = .70) in previous studies (Xue et al, 2017).

Brief-Spanish English Verbal Learning Test.—The Spanish English Verbal Learning Test (González, Mungas, Reed, Marshall, & Haan, 2001) is a 15-item verbal learning and recall task developed to be psychometrically equivalent across English and Spanish translations. The Spanish English Verbal Learning Test was developed by generating an English list of items commonly found in grocery stores, translating items to Spanish, and back-translating to English. In this task, participants are presented the 15-item list across five learning trials, with words presented at a rate of one-and-a-half seconds per word. After each learning trial, participants are asked to repeat as many words as they can recall within 60 seconds. Following the learning trials, a distraction list of 15 words is presented, and the participant is asked to repeat each word as they are presented. Immediately after the presentation of the distractor list, the participants are asked to recall the first set of 15 words within 60 seconds. In comparison to the original Spanish English Verbal Learning Test which had 5 learning trials, four Brief-Spanish English Verbal Learning Test (B-SEVLT) scores are derived, one for each of the three learning trials as well as one for the delayed recall trial (0–15 per trial).

Word Fluency.—Word Fluency (WF) is a task of verbal fluency derived from the Multilingual Aphasia Examination during which participants are asked to produce as many words as possible beginning with the letters F and A, within a 60-second period per letter. In the HCHS/SOL, responses in English and Spanish were permissible while proper nouns, locations, and numbers were not allowed. WF assesses vocabulary, verbal processing speed, retrieval, and response inhibition (Shao, Janse, Visser, & Meyer, 2014). Verbal fluency tasks have a long history in neuropsychology and are included in many neuropsychology tests as measures of flexibility and processing speed, such as the Delis-Kaplan Executive Functioning System (D-KEFS; Delis, Kaplan, & Kramer, 2001).

Digit Symbol Substitution.—The Digit Symbol Substitution test (DSS), a measure of processing speed and attention that has been utilized in neurocognitive testing for over a century (Jaeger, 2018), was derived from the Wechsler Adult Intelligence Scale – Revised (WAIS-R). Participants are required to transpose digits (from 1 to 9) into symbols based on a key in a 90-second period. Scores are calculated as the total number of correct substitutions completed within the time limit.

Procedures

The institutional review boards at each site approved the study protocols and procedures. Participants provided informed consent prior to participation. Participants of the baseline HCHS/SOL age 45 years and older were administered a brief neurocognitive battery consisting of the SIS, B-SEVLT, WF, and DSS, in that fixed order. Testing at each field center occurred in a quiet area to minimize distraction. Testing was conducted by examiners trained to proficiency and certified by field center lead examiners. Measures were administered in the examinee's preferred language. Participants were allowed to wear reading glasses or hearing aids. All tests were administered in a standardized fashion, with instruction scripts read aloud exactly as written by examiners. To establish Spanish versions of the SIS, WF, and DSS, English versions were translated into Spanish and back translated to English. Translations and back translations were evaluated by the Translation & Validation subcommittee of the HCHS/SOL to ensure equivalence across background groups (Sorlie et al., 2010).

Statistical Analyses

Confirmatory factor analysis (CFA) was used to test the fit of a single factor model for each neuropsychological instrument as well as a second-order factor model comprised of the SIS, B-SEVLT, and WF first-order factors in addition to the single item DSS within each language group. Raw scores were used for all observed indicators. *Mplus* version 8.3 (Muthen & Muthen, 2019) was used to estimate the models. Multiple-group CFA was then used to test three nested levels of measurement invariance across each language of administration at the item-level as well as at the total score level. Differences between levels of invariance were evaluated by comparing the chi-square statistic (χ^2) of the most constrained model to the prior model. Specifically, a chi-square difference ($\Delta\chi^2$) test between nested models subtracts the more constrained model's chi-square statistic from that of prior models with the difference in the models' degrees of freedom. A significant chi-square difference test indicates the model fit is significantly worse in the constrained model

compared to the previous model. DIFFTEST in Mplus, a corrected chi-square difference test, was used to calculate difference tests of models estimated with weighted least squares mean and variance (WLSMV) estimation, while R was used to calculate difference tests of models estimated with maximum likelihood (ML) estimation. In addition to significant chi-square difference tests, decreases in CFI ≥ .01 and increases in RMSEA ≥ .01 suggested meaningful decreases in fit due to equality constraints (Chen, 2007).

First, configural invariance was tested by constructing each factor model in both English and Spanish groups without equality constraints and confirming the models fit adequately. Metric invariance was then assessed by constraining factor loadings across language groups. Similarly, scalar invariance was tested by maintaining equality constraints in factor loadings that were found to be invariant and also constraining item thresholds/intercepts. In cases of measurement non-invariance, partial measurement invariance was explored by testing for the parameters that produced the greatest misfit when constrained. Each parameter constrained in the non-invariant model was freed one-at-a-time and compared to the previous, invariant model. Non-significant change in fit between these models support the presence of partial invariance. In the case of partial metric invariance, the item with non-invariant factor loadings would not have its threshold/intercept constrained in the scalar model (Thompson & Green, 2013). Standardized factor loadings (λ) for each configural model are reported in Table 2.

Given metric and scalar invariance, differences in the factor means can be compared between groups by estimating the factor mean of the Spanish-speaking group while setting the mean of the English-speaking group to zero. A significant benefit of comparing factor means between groups in this manner is the ability to identify and model non-invariance in parameters across groups (Thompson & Green, 2013). By modeling partial invariance in factor loadings or item thresholds/intercepts, the group means can still be compared within the CFA framework; the presence of non-invariance would invalidate the use of raw sum or composite scores to compare group means, as is often done within the general linear model (GLM) framework (i.e., regression & ANOVA).

Results

Six-Item Screener

Weighted least squares mean and variance (WLSMV) estimation, appropriate for dichotomous items, was used in a CFA of the SIS. Model fit indices for all invariance models tested are reported in Table 3. The configural model, without parameter constrains across groups, did not demonstrate good fit across all indices; however, once a covariance between the residuals of SIS items 2 (“What month is this?”) and 3 (“What is the day of the week?”) was specified, model fit was acceptable. Constraining the factor loadings (i.e., metric model) did not considerably change fit indices ($\Delta\chi^2$ (5) = 7.07, p = .216). However, constraining item thresholds (i.e., scalar model) significantly decreased model fit ($\Delta\chi^2$ (5) = 18.05, p = .003) and the decrease was meaningful (ΔCFI = .04, ΔRMSEA = .01).

To examine which item(s) demonstrated significant non-invariance in difficulty across language, each item threshold was freed one-at-a-time. The model that allowed the threshold

of item 6 to vary across groups demonstrated the best fit, and was not significantly different from the metric model ($\Delta\chi^2$ (4) = 3.46, p = .484). The threshold for item 6 ("Sofa") in the metric model was greater in the Spanish group (γ = −0.76, *S.E.* = 0.11, p < .001) compared to the English-speaking group (γ = −1.14, *S.E.* = 0.09, p < .001), indicating Spanish-speaking participants with the same neurocognitive ability were more likely to forget this item compared to English-speaking participants. Finally, difference in neurocognitive impairment across language of administration groups was tested by examining the significance of the group difference on the latent variable (M = −0.24, *S.E.* = 0.11, p = .020), indicating that participants who responded in Spanish demonstrated greater neurocognitive impairment than those who responded in English, after matching participants on key demographic variables and accounting for partial scalar non-invariance.

Brief-Spanish English Verbal Learning Task

Similar CFA procedures were conducted on the B-SEVLT to test for measurement invariance across languages. As scores on the B-SEVLT items ranged from 0 to 15 and were approximately normally distributed, maximum likelihood estimation (ML) was used. The fit indices of the configural model were mixed. Similarly, the fit indices of the model with constrained factor loadings were mixed, although the fit was not significantly worse than the configural model ($\Delta\chi^2$ (3) = 5.86, p = .119). Constraining item intercepts did not significantly worsen fit either ($\Delta\chi^2$ (3) = 5.42, p = .144), and overall model fit was acceptable. Lastly, the mean of the Spanish-speaking group was significant (M = −0.25, *S.E.* = 0.06, p < .001), suggesting participants who responded in Spanish had lower verbal learning and recall ability than English respondents.

Word Fluency

As WF in the HCHS/SOL only included two items (letters F and A), measurement invariance could not be tested for WF alone. Therefore, invariance of WF was embedded in a correlated-factors model that included SIS and B-SEVLT latent variables. WLSMV was used as the items of the SIS are dichotomous. Correlations between SIS, B-SEVLT, and WF latent variables were freely estimated. Due to the previously reported non-invariance, means in SIS and B-SEVLT were freed in the Spanish group, as was the threshold of SIS item 6.

The configural model fit well, as did the metric model, and these models were not significantly different ($\Delta\chi^2$ (1) = 1.26, p = .261). However, constraining item intercepts significantly decreased model fit ($\Delta\chi^2$ (1) = 102.82, p < .001) and the decrease was meaningful (ΔCFI = .01, ΔRMSEA = .01). The intercept of letter A was lower in the English group (γ = 9.14, *S.E.* = 0.14, p < .001) compared to the Spanish group (γ = 10.97, *S.E.* = 0.20, p < .001), suggesting that given the same level of verbal fluency, Spanish participants will perform better on this item compared to English participants. However, the mean of overall WF was significantly lower in the Spanish group (M = −1.55, *S.E.* = 0.20, p < .001), indicating those who were assessed in Spanish demonstrated lower verbal fluency than English respondents.

Neurocognitive Functioning

The final set of CFA models tested were second-order latent variable models of neurocognitive functioning. The neurocognitive functioning latent variable was indicated by the SIS, B-SEVLT, and WF latent variables, in addition to DSS scores, and was estimated with WLSMV. In the configural model, the means of the SIS, B-SEVLT, and WF were allowed to vary in the Spanish group, as were the intercepts of SIS item 6 and WF letter A. The conceptual model is depicted in Figure 1.

The configural model fit well, as did the metric model; however, the fit was significantly worse after constraining factor loadings ($\Delta\chi^2$ (3) = 11.03, p = .012), although the decrease in the fit was marginal (ΔCFI < .01, ΔRMSEA = .01). To test for partial metric invariance, each factor loading was freed one-at-a-time and model fit was compared to the configural model. Ultimately, freeing the loading of WF on neurocognitive functioning improved model fit, which was not significantly worse than the configural model ($\Delta\chi^2$ (2) = 5.15, p = .076). The relationship between WF and neurocognitive functioning was somewhat stronger in the Spanish group (b = 6.42, *S.E.* = 1.86, p = .001, β = .72) than in the English group (b = 4.83, *S.E.* = 1.46, p = .001, β = .57).

Scalar invariance of neurocognitive functioning was tested by constraining the intercepts of the SIS, B-SEVLT, and DSS; WF was unconstrained given the partial metric non-invariance. The model demonstrated acceptable fit, but fit significantly worse than the partial metric model ($\Delta\chi^2$ (2) = 45.86, p < .001, ΔCFI = .01, ΔRMSEA = .01). Again, each intercept was freed one-at-a-time and model fit was compared to the partial metric model. Ultimately, freeing the intercept of DSS resulted in the best fitting model, which was not significantly different from the partial metric model ($\Delta\chi^2$ (1) = 0.36, p = .548). When unconstrained, the intercept of DSS was greater in the English-speaking group (γ = 44.42, *S.E.* = 0.42, p < .001) compared to in the Spanish-speaking group (γ = 39.16, *S.E.* = 0.66, p < .001). These results indicate that English language test-takers completed five more substitutions on average, suggesting higher processing speed than Spanish language test-takers at the same overall neurocognitive ability. Lastly, in this final model, the mean of neurocognitive functioning was significantly lower in the Spanish group (M = −0.10, *S.E.* = 0.01, p = .001). Taken together, after accounting for differential functioning at the item and first-order latent variable levels, participants who responded in Spanish demonstrated significantly lower broad neurocognitive functioning than those who responded in English, regardless of Hispanic/Latino background.

Discussion

This study aimed to tease apart differences in the performance of a brief neurocognitive battery across Spanish language and English language administrations. We explored potential differential functioning at both item and factor levels through a series of CFA models. Overall, the neurocognitive measures included in the HCHS/SOL were partially invariant across Spanish and English administrations at the item level. Moreover, the B-SEVLT demonstrated item-level invariance at the configural, metric, and scalar levels, demonstrating the psychometric value of measured developed in Spanish and English simultaneously; however, Spanish test-takers performed significantly worse on the B-SEVLT

at the factor-level, suggesting some language-based differences may remain. There was also some evidence of bias in item thresholds/intercepts on the SIS and WF, revealing that the difficulty of these assessments may not be consistent across languages despite the fact that groups were matched across age, sex, education, immigration status, and Hispanic/Latino background. Similar difference in intercepts was indicated at the broader neurocognitive functioning level, with Spanish-speaking test-takers performing worse on the DSS task.

Non-invariance in the loading of WF on neurocognitive functioning revealed differences in the relevance of verbal fluency and vocabulary to overall neurocognitive functioning when assessed in English versus Spanish. For Spanish participants, the strength of the relationship between WF and neurocognitive functioning was comparable to other neurocognitive assessments, whereas the relationships was significantly weaker for English participants. Metric non-invariance can be a threat to test validity as it suggests the construct may have a different meaning across groups – in this case, verbal fluency was more strongly related to neurocognitive functioning in Spanish-speaking participants than in their English-speaking counterparts.

Additionally, English-speaking respondents performed significantly better on the DSS than their Spanish-speaking counterparts when controlling for overall neurocognitive ability, translating to approximately five more digits on average. Previous work has suggested aspects of digit-based neuropsychological assessments may be more difficult in Spanish (López, Steiner, Hardy, IsHak, & Anderson, 2016), perhaps due to linguistic differences in the syllables and reading rate associated with each digit (Naveh-Benjamin & Ayres, 1986). While these previous analyses have focused on the interaction between language and digits in regards to working memory, the same mechanisms may affect the neurocognitive processing speed of reading and translating digits during the DSS task. Differential performance between U.S. American and South American participants on timed, but not untimed, processing speed tasks (Cores et al., 2015) supports this hypothesis.

Perhaps more striking is the consistent difference in latent variable means across first-order factors (i.e., SIS, B-SEVLT, & WF) as well as the second-order neurocognitive functioning factor across language administration. Test-takers in Spanish consistently performed worse than test-takers in English across all neuropsychological measures. Given that participants were matched on age, gender, years of education, and US immigration status, this consistent difference may suggest that some amount of bias exists in these measurements above and beyond the variables matched for, including additional confounding variables previously identified in cultural neuropsychology (Ramirez et al., 2006). For example, differences in the structure of language may lend itself to tasks of verbal learning and fluency, depending on factors such as differential exposure to specific words, word length, and the number of syllables in words (Kempler, Teng, Dick, Taussig, & Davis, 1998), which can manifest as higher performance on verbal fluency tasks in one language over another. Further, this study demonstrates that both measures developed in English and translated to Spanish (e.g., SIS, WF, & DSS), as well as measures developed in Spanish and English simultaneously (e.g., B-SEVLT), are susceptible to measurement inequivalence. This is consistent with past measurement invariance studies that have indicated tests developed in Spanish and English simultaneously can still demonstrate language-based biases (Mungas et al., 2000; Mungas

et al., 2011), although simultaneous development with measurement invariance procedures does allow for language validation during test development. Taken together, neither accurate translation of neuropsychological measures from English to Spanish nor simultaneous development in Spanish and English are sufficient to ensure psychometric equivalence, and measurement invariance procedures remain necessary to identify and eliminate test biases.

Neuropsychologists have become increasingly aware of the influence that broad cultural factors can have on neurocognitive performance and warn that neuropsychological assessments may not perform equivalently across cultural groups (Fernandez & Marcopulos, 2019). While the current study controlled for several of the obvious potential confounding variables (e.g., age, education, & nationality), there are several other cultural variables that were not included that may help to account for the consistent underperformance of Spanish-speaking test-takers rather than pure item or test bias. For example, previous research has suggested acculturation to the mainstream US culture is associated with decreased neuropsychological performance on verbal fluency and processing speed (Boone, Victor, Wen, Razani, & Pontón, 2007). On the other hand, there is some evidence that acculturation is related to better neuropsychological performance (Tan, Burgess, & Green, 2020), although relationships are inconsistent and effect sizes were often small. Additionally, differences in cultural attitudes or perceptions of timed tests (Cores et al., 2015) may contribute to some of the differences observed here. Given this evidence, clinicians and researchers working with diverse populations should be cognizant of how cultural attitudes and experiences may influence neuropsychological testing when considering timed versus untimed tests. While this explanation would not suffice for untimed assessments such as the SIS, the broader "approach to standardized testing," which has been identified as a potential cross-cultural confound, may help explain differences in performance on untimed assessments. The present study reveals that while some degree of item-level bias in these neuropsychological measures is likely attributable to the language of assessment, other cultural factors may be responsible for the remaining, systematic underperformance of Spanish-speaking examinees, which we review below. Further research that parses apart the differential performance of Spanish-speaking and English-speaking test-takers by investigating cultural influences is certainly warranted.

One cultural and linguistic influence that may contribute to these findings is the extent to which bilingualism affects performance on neuropsychological measures. Lamar and colleagues (2019) demonstrated self-reported bilingual proficiency is positively associated with cognitive performance in Hispanics/Latinos. Additionally, bilinguals with a dominant language demonstrated significant differences in neuropsychological performance based on the language of assessment, whereas bilinguals with equal proficiency in Spanish and English did not (Gasquoine, Croyle, Cavazos-Gonzalez, & Sandoval, 2007). On the other hand, other work has suggested there are not differences in processing speed, verbal fluency, and lexical access between Hispanics/Latinos who learned English as a primary or secondary language (Boone, Victor, Wen, Razani, & Pontón, 2007). Further, the small body of literature examining the influence of the language of assessment on the neuropsychological performance of bilinguals has suggested that those who undergo testing in Spanish perform worse than both bilinguals and monolinguals assessed in English (Ott et al., 2014). Consideration of bilingualism and language fluency may add additional

layers of complexity when considering the effect of language on neuropsychological performance (Mindt et al., 2008). Those findings in conjunction with the present study suggest that cultural and linguistic influences above and beyond bilingualism may affect the psychometric properties of neuropsychological measures, leading to underperformance of Spanish examinees.

Another consideration that may help to explain our results is that English test-takers in this sample could possess higher neurocognitive functioning secondary to bilingualism or to occupations that require greater proficiency in English. Thus, Hispanic/Latino individuals who fit into these categories may be more likely to adopt English as a primary or preferred language, and thus undergo examination in English. Additionally, language may serve as a proxy for socioeconomic status not otherwise controlled for in these analyses. Those who underwent assessment in English may also have better access to health care and preventative medicine, which may in turn result in somewhat higher neurocognitive functioning in adulthood. At the current level of cross-sectional analysis, it would not be possible to disentangle if current neuropsychological performance is a consequence of inherent test bias or rather a byproduct of greater neurocognitive ability also affecting language development, bilingualism, acculturation, or occupational or socioeconomic status. Consequently, further longitudinal psychometric analyses which can control for the influence of these variables over time would be necessary to parse these possible interactions.

A second potential source of cultural influence is that most neuropsychological assessments are developed in English and subsequently translated (Siedlecki et al., 2010). The results of this study support previous observations that accurate translation may not sufficiently account for the impact linguistic and cultural differences can have on an assessment's psychometric properties (Mungas et al., 2000). Additionally, while neurocognitive assessments of the HCHS/SOL were translated and back translated and evaluated by committee for accuracy to establish linguistic equivalence across languages, subtle linguistic deviations which affect performance may persist. For example, regional differences within the Spanish language may contribute to inequivalence in psychometric properties (González et al., 2019), including item and test difficulty. Taken together, even accurate and validated translational techniques may not fully capture cross-cultural nuances in neuropsychological instruments. Consequently, assessments may need to be more constructed specifically in the language in which they are utilized. To this point, the Multilingual Aphasia Examination (MAE), an instrument adapted to Spanish, but not directly translated, has demonstrated equivalent performance at the scale-level in Spanish and English (Rey, Feldman, Rivas-Vazquez, Levin, & Benton, 1999); however, item-level equivalence was not examined. In the case of phonemic fluency, the English and Spanish version of the MAE use different letters, which prevents item-level invariance across test forms. By utilizing the same letters across language, the present study was able to test violations of invariance at the item level.

The motivation for conducting this study was to explore the psychometric properties of neurocognitive instruments across language as one avenue to explain differences in Hispanic/Latino heritage groups previously identified (e.g., Buré-Reyes et al., 2013; González et al., 2015). While these findings demonstrated that English examinees performed higher than Spanish test-takers on each domain given the same neurocognitive ability,

our findings do not necessarily indicate that this disparate performance is the sole result of systematic bias in these neuropsychological assessments. If this were the case, more pervasive differential functioning at the item level would have been observed across languages. Therefore, while the language of assessment may have an influence on neuropsychological performance, the previously identified differences within the greater Hispanic/Latino population is likely a consequence of a combination of factors including additional cultural factors and/or individual differences not measured in this study. Thus, it is possible that the language of test development has an unintentional influence on the psychometrics of the test that is not fully accounted for by translation or norming procedures (Artiola i Fortuny & Mullaney, 1997). Further empirical examinations of the psychometric properties of neuropsychological tests across language and culture is warranted to fully understand what may be driving group differences.

There are some limitations of note in the present study. While several potentially confounding demographic variables were controlled for (i.e., age, years of education, immigration status), it was not possible to control for the *quality* of education across groups, which may serve as an alternate explanation to the observed findings. It has been previously revealed that ethnic and cultural differences in neuropsychological performance are sensitive to education quality above and beyond demographic characteristics (Fyffe et al., 2011; Manly, Jacobs, Touradji, Small, & Stern, 2002), which may help explain the consistent difference in performance observed in this study. The use of norms which are sensitive to sociocultural and demographic factors, including language of assessment, years in the U.S., educational attainment, language of education, and heritage group are necessary for within-group comparisons; however, implementing group-specific norms will not address measurement non-invariance, which serves to bias between-group comparisons. Stopping at the development of norms specific to ethnic, cultural, or linguistic group fails to address deeper inequivalences in the measurement and construct validity of neuropsychological domains developed in well-educated, Caucasian samples, which may be less generalizable to diverse populations (Manly, 2005).

There are additional limitations surrounding linguistic factors as well. Participants chose whether to undergo assessment in English or Spanish, and a formal language assessment was not used to determine the language of testing. It may be the case that participants' preferred language differed from their most proficient language, resulting in differences in neuropsychological performance, although some evidence suggests language preference and proficiency are strongly related (Gee, Walsemann, & Takeuchi, 2010). Some assessments did allow responding in Spanish or English regardless of the language of administration, but the discrepancy between language preference and language proficiency may contribute to some of these findings. Bias in the administration of testing across groups, a specific form of method bias, was not assessed in this study whereas previous research has indicated method biases can have a tangible influence on neuropsychological testing in cross-cultural contexts (Fernández & Abe, 2018). Differences in the interactions between administrators and participants, such as dialectal differences or participant experiences during testing, can negatively impact performance (Fernández & Abe, 2018), and these factors were not accounted for. Lastly, this study included both timed and untimed assessments. Timed tests of processing speed, such as the DSS, may be more sensitive to cross-cultural bias due

to differences in cultural attitudes and familiarity (Cores et al., 2015), and this bias may contribute to the differences across language observed in this study. Further work should clarify if language-based differences in performance persist in timed and untimed processing speed assessments.

Despite these limitations, this study provides preliminary evidence differences in the language of assessment may affect the psychometric properties of neuropsychological assessments. Thus, continued attention must be paid to the structure of assessments in diverse populations. Further, merely assuming measurement equivalence, even of instruments developed simultaneously across languages, is not sufficient, and possible measurement discrepancies should be investigated if possible, or at least considered when interpreting results. While the development of norms may attempt to correct for differences arising from non-invariance (Pedraza & Mungas, 2008), merely adjusting norms does not sufficiently address ethnic, cultural, or linguistic biases (Manly, 2005). Therefore, continued development of instruments with equivalent psychometric properties across ethnic, cultural, and linguistic populations of interest, for both clinical and research purposes, is paramount in cultural neuropsychology.

Acknowledgements

The authors thank the staff and participants of HCHS/SOL for their important contributions. A complete list of staff and investigators has been provided by Sorlie et al. (2010) and is also available on the study website: http://www.cscc.unc.edu/hchs/

Funding

Zachary T. Goodman is supported by the National Institutes of Health (T32-HL007426). The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements.

References

Artiola i Fortuny L, & Mullaney HA (1997). Neuropsychology with Spanish speakers: Language use and proficiency issues for test development. Journal of Clinical and Experimental Neuropsychology, 19(4), 615–622. 10.1080/01688639708403747 [PubMed: 9342693]

Bezdicek O, Moták L, Schretlen DJ, Preiss M, Axelrod BN, Nikolai T … Růžička E (2016). Sociocultural and language differences on the Trail Making Test. Archives of Assessment Psychology, 6(1), 33–48.

Brewster PWH, Melrose RJ, Marquine MJ, Johnson JK, Napoles A, MacKay-Brandt A, … Mungas D (2014). Life experience and demographic influences on cognitive function in older adults. Neuropsychology, 28(6), 846–858. 10.1037/neu0000098 [PubMed: 24933483]

Boone KB, Victor TL, Wen J, Razani J, & Pontón M (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. Archives of Clinical Neuropsychology, 22(3), 355–365. 10.1016/j.acn.2007.01.010 [PubMed: 17320344]

Buré-Reyes A, Hidalgo-Ruzzante N, Vilar-López R, Gontier J, Sánchez L, Pérez-García M, & Puente AE (2013). Neuropsychological test performance of Spanish speakers: Is performance

different across different Spanish-speaking subgroups? Journal of Clinical and Experimental Neuropsychology, 35(4), 404–412. 10.1080/13803395.2013.778232 [PubMed: 23496164]

Byrne BM, Shavelson RJ, & Muthén B (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105(3), 456–466. 10.1037/0033-2909.105.3.456

Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, & Hendrie HC (2002). Six-Item Screener to Identify Cognitive Impairment among Potential Subjects for Clinical. Care, 40(9), 771–781. 10.1097/01.MLR.0000024610.33213.C8

Chen FF (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14(3), 464–504. 10.1080/10705510701301834

Cores EV, Vanotti S, Eizaguirre B, Fiorentini L, Garcea O, Benedict RHB, & Cáceres F (2015). The effect of culture on two information-processing speed tests. Applied Neuropsychology: Adult, 22(4), 241–245. 10.1080/23279095.2014.910214 [PubMed: 25372116]

Delis DC.; Kaplan E; Kramer JH (2001). Delis-Kaplan Executive Function System: Technical Manual. San Antonio, TX: Harcourt Assessment Company.

Fernández AL, & Abe J (2018). Bias in cross-cultural neuropsychological testing: problems and possible solutions. Culture and Brain, 6(1), 1–35. 10.1007/s40167-017-0050-2

Fyffe DC, Mukherjee S, Barnes LL, Manly JJ, Bennett DA, & Crane PK (2011). Explaining differences in episodic memory performance among older African Americans and Whites: The roles of factors related to cognitive reserve and test bias. Journal of the International Neuropsychological Society, 17(4), 625–638. 10.1017/S1355617711000476 [PubMed: 23131601]

Gasquoine PG, Croyle KL, Cavazos-Gonzalez C, & Sandoval O (2007). Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults. Archives of Clinical Neuropsychology, 22, 991–1001. doi:10.1016/j.acn.2007.08.003 [PubMed: 17900857]

González HM, Mungas DAN, Reed BR, Marshall S, & Haan MN (2001). A new verbal learning and memory test for English- and Spanish-speaking older people. Journal of the International Neuropsychological Society, 7(5), 544–555. 10.1017/S1355617701755026 [PubMed: 11459106]

González HM, Tarraf W, Fornage M, González KA, Chai A, Youngblood M, … Schneiderman N (2019). A research framework for cognitive aging and Alzheimer's disease among diverse US Latinos: Design and implementation of the Hispanic Community Health Study/Study of Latinos —Investigation of Neurocognitive Aging (SOL-INCA). Alzheimers' & Dementia, 15(12), 1624–1632. 10.1016/j.jalz.2019.08.192

González HM, Tarraf W, Gouskova N, Gallo LC, Penedo FJ, Davis SM, … Mosley TH (2015). Neurocognitive function among middle-aged and older Hispanic/Latinos: Results from the Hispanic Community Health Study/Study of Latinos. Archives of Clinical Neuropsychology, 30(1), 68–77. 10.1093/arclin/acu066 [PubMed: 25451561]

Hedden T, Park DC, Nisbett R, Ji LJ, Jing Q, & Jiao S (2002). Cultural variation in verbal versus spatial neuropsychological function across the life span. Neuropsychology, 16(1), 65–73. 10.1037/0894-4105.16.1.65 [PubMed: 11853358]

Helms JE (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47(9), 1083–1101. doi:10.1037/0003-066x.47.9.1083

Jaeger J (2018). Digit symbol substitution test. Journal of Clinical Psychopharmacology, 38(5), 513–519. 10.1097/JCP.0000000000000941 [PubMed: 30124583]

Jones RN (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care, 44(1), 124–133. 10.1097/01.mlr.0000245250.50114.0f [PubMed: 16434911]

López E, Steiner AJ, Hardy DJ, IsHak WW, & Anderson WB (2016). Discrepancies between bilinguals' performance on the Spanish and English versions of the WAIS Digit Span task: Cross-cultural implications. Applied Neuropsychology:Adult, 23(5), 343–352. 10.1080/23279095.2015.1074577 [PubMed: 26786894]

Fernandez AL & Marcopulos BA (2018). Cross-cultural tests in neuropsychology. A review of recent studies and a modest proposal. In Koffler S, Mahone EM, Marcopulos BA, Johnson-Greene DE, & Smith G (Eds.) Neuropsychology: Science and Practice, III (pp. 93). Oxford.

Folstein MF, Folstein SE, & McHugh PR (1975). "Mini-mental status". A practical method for grading the cognitive state of patients for the clinician". Journal of Psychiatric Research, 12(3), 189–198. doi:10.1016/0022-3956(75)90026-6. [PubMed: 1202204]

Lamar M, León A, Romo K, Durazo-Arvizu RA, Sachdeva S, … González HM (2019). The independent and interactive associations of bilingualism and sex on cognitive performance in Hispanics/Latinos of the Hispanic Community Health Study/Study of Latinos. Journal of Alzheimer's Disease, 71(4), 1271–1283. doi: 10.3233/JAD-190019.

LaVange LM, Kalsbeek W, Sorlie PD, Avilés-Santa LM, Kaplan RC … Elder JP (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Annals of Epidemiology, 20(8), 642–649. 10.1016/j.annepidem.2010.05.006 [PubMed: 20609344]

Manly JJ (2005). Advantages and disadvantages of separate norms for African Americans. Clinical Neuropsychologist. 10.1080/13854040590945346

Manly JJ, Jacobs DM, Touradji P, Small SA, & Stern Y (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. Journal of the International Neuropsychological Society, 8(3), 341–348. 10.1017/S1355617702813157 [PubMed: 11939693]

Meredith W (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525–543.

Meredith W, & Teresi JA (2006). An essay on measurement and factorial invariance. Medical Care, 44(11 SUPPL. 3), 69–77. 10.1097/01.mlr.0000245438.73837.89

Mindt MR, Arentoft A, Germano KK, D'Aquila E, Scheiner D, Pizzirussom M, Sandoval TC, & Gollan TH (2008). Neuropsychological, cognitive, and theoretical considerations for evaluation of bilingual individuals. Neuropsychological Review, 18(3), 255–268. doi:10.1007/s11065-008-9069-7.

Mungas D, Reed BR, Haan MN, González HM (2005). Spanish and English neuropsychological assessment scales: relationship to demographics, language, cognition, and independent function. Neuropsychology, 19(4), 466–475. [PubMed: 16060821]

Mungas D, Reed BR, Sarah Marshall MC, & González HM (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology, 14(2), 209–223. 10.1037/0894-4I05.14.2.209 [PubMed: 10791861]

Mungas D, Widaman KF, Reed BR, & Tomaszewski Farias S (2011). Measurement invariance of neuropsychological tests in diverse older persons. Neuropsychology, 25(2), 260–269. 10.1037/a0021090 [PubMed: 21381830]

Mungas D, Reed BR, Crane PK, Haan MN, González H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. Psychol Assess.16(4), 347–359. [PubMed: 15584794]

Muthén LK, & Muthén BO (2019). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Naveh-Benjamin M, & Ayres TJ (1986). Digit Span, Reading Rate, and Linguistic Relativity. The Quarterly Journal of Experimental Psychology Section A, 38(4), 739–751. 10.1080/14640748608401623

Ott S, Schatz P, Solomon G, & Ryan JJ (2014). Neurocognitive performance and symptom profiles of Spanish-speaking Hispanic athletes on the impact test. Archives of Clinical Neuropsychology, 29(2), 152–163. 10.1093/arclin/act091 [PubMed: 24389704]

Pedraza O, & Mungas D (2008, 9). Measurement in cross-cultural neuropsychology. Neuropsychology Review. 10.1007/s11065-008-9067-9

Puente AE, & Perez-Garcia M (2000). Chapter 20 – Neuropsychological Assessment of Ethnic Minorities: Clinical Issues. In Handbook of Multicultural Mental Health. 10.1016/B978-012199370-2/50021-3

Ramirez-Zohfeld V, Rademaker AW, Dolan NC, Ferreira MR, Eder MM, Liu D, … Cameron KA (2015). Comparing the Performance of the S-TOFHLA and NVS among and between English and Spanish Speakers. Journal of Health Communication. 10.1080/10810730.2015.1018629

Rey GJ, Feldman E, Rivas-Vazquez R, Levin BE, & Benton A (1999). Neuropsychological test development and normative data on Hispanics. Archives of Clinical Neuropsychology, 14(7), 593–601. 10.1016/S0887-6177(99)00008-6 [PubMed: 14590573]

Shao Z, Janse E, Visser K, & Meyer AS (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. Frontiers in Psychology, 5, 1–10. 10.3389/fpsyg.2014.00772 [PubMed: 24474945]

Siedlecki KL, Manly JJ, Brickman AM, Schupf N, Tang MX, & Stern Y (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? Neuropsychology, 24(3), 402–411. 10.1037/a0017515 [PubMed: 20438217]

Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, … & LaVange L (2010). Design and implementation of the Hispanic Community Health Study/Study of Latinos. Annals of Epidemiology, 20(8), 629–641. [PubMed: 20609343]

Tan YW, Burgess GH, & Green RJ (2020). The effects of acculturation on neuropsychological test performance: A systematic literature review. The Clinical Neuropsychologist, 1–31. doi:10.1080/13854046.2020.1714740

Thompson MS, & Green SB (2013). Evaluating between-group differences in latent variable means. In Structural Equation Modeling: A Second Course. Hancock GR, & Mueller RO (Second Edition). 163–218. Greenwich, Conn: IAP.

Tuokko HA, Chou PHB, Bowden SC, Simard M, Ska B, & Crossley M (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. Journal of the International Neuropsychological Society, 15(3), 416–425. 10.1017/S1355617709090602 [PubMed: 19402928]

U.S. Census Bureau. (2010). The older population in the United States: 2010 to 2050. 10–33. https://www.census.gov/content/dam/Census/library/publications/2010/demo/p25-1138.pdf

Wechsler DWechsler Adult Intelligence Scale–Fourth Edition. Pearson; San Antonio, TX: 2008.

Xue J, Chiu HFK, Liang J, Zhu T, Jiang Y, & Chen S (2018). Validation of the Six-Item Screener to screen for cognitive impairment in primary care settings in China. Aging and Mental Health, 22(4), 453–457. 10.1080/13607863.2017.1280768 [PubMed: 28145741]

Zubizarreta JR, Paredes RD, & Rosenbaum PR (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. Annals of Applied Statistics, 8(1), 204–231. 10.1214/13-AOAS713

Key Points

Question:

Does a brief neurocognitive assessment provide an equivalent estimate of neurocognitive abilities if administered in English or Spanish to Hispanic/Latino adults in the United States?

Findings:

Test takers who underwent assessment in Spanish performed worse than those who were assessed in English, demonstrating poorer neurocognitive functioning despite controlling for several sociodemographic factors.

Importance:

This study highlights how linguistic differences can influence the properties of assessments developed in English and translated to Spanish, and the impact those discrepancies have on estimating neuropsychological abilities.

Next Steps:

Further research is necessary to elucidate sources of bias which contribute to differential performance across language and cultural groups, and how these biases influence performance in clinical and research settings.

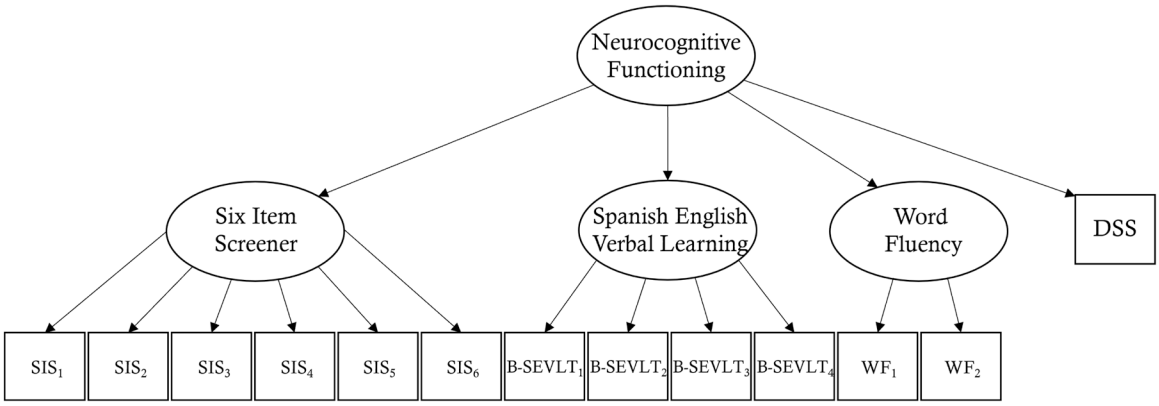

Figure 1.

Conceptual representation of the second-order neurocognitive CFA model before the inclusion of residual covariance between two items of the SIS.

Table 1

Demographic characteristics of the Spanish and English-speaking Hispanic/Latino samples in the HCHS/SOL, with statistical comparisons between groups.

	Spanish		English		
Variable	***M***	**(SD)**	***M***	**(SD)**	**Statistical Test**
Age	54.18	(7.19)	53.64	(6.64)	$t(1695.3) = 1.62$, $p = .105$
Years of Education	12.49	(4.71)	12.80	(3.18)	$t(1497.5) = 1.61$, $p = .107$
	n	(%)	*n*	(%)	
Women	515	(60.3)	515	(60.3)	$\chi^2(1) = 0.00$, $p = 1.00$
US-born	657	(76.9)	657	(76.9)	$\chi^2(1) = 0.00$, $p = 1.00$
Heritage Group					
Puerto Rican	557	(65.5)	557	(65.5)	
Mexican	202	(23.8)	202	(23.8)	
Cuban	39	(4.6)	39	(4.6)	$\chi^2(1) = 0.00$, $p = 1.00$
Dominican	21	(2.5)	21	(2.5)	
Central American	18	(2.1)	18	(2.1)	
South American	17	(2.0)	17	(2.0)	

Note. U.S.-born includes territories (e.g., Puerto Rico).

Table 2

Standardized factor loadings from the respective configural models of each factor.

Variable	English-Speakers	Spanish-Speakers
SIS		
SIS_1	.66	.28
SIS_2	.27	.44
SIS_3	.47	.47
SIS_4	.79	.69
SIS_5	.77	.67
SIS_6	.55	.54
B-SEVLT		
$B\text{-}SEVLT_1$	.69	.65
$B\text{-}SEVLT_2$	.85	.86
$B\text{-}SEVLT_3$	.88	.87
$B\text{-}SEVLT_4$	.82	.82
WF		
WF_1	.71	.81
WF_2	.82	.81
NF		
SIS	.76	.70
B-SEVLT	.69	.72
WF	.73	.57
DSS	.71	.63

Note. SIS = Six-Item Screener; B-SEVLT = Brief-Spanish English Verbal Learning Test; WF = Word Fluency; DSS = Digit Symbol Substitution; NF = Neurocognitive Functioning.

Table 3

Fit indices for all measurement invariance CFA models tested.

Factor	χ^2	*df*	*p*	CFI	RMSEA	SRMR
SIS						
Configural$_a$	44.15	18	< .001	.91	.04 [.03, .06]	.10
Configural$_b$	25.94	16	.055	.97	.03 [.00, .05]	.07
Metric	31.61	21	.064	.96	.02 [.00, .04]	.09
Scalar	48.88	26	.004	.92	.03 [.02, .05]	.09
Partial Scalar	35.11	25	.086	.96	.02 [.00, .04]	.09
B-SEVLT						
Configural	52.87	4	< .001	.99	.12 [.09, .15]	.02
Metric	58.73	7	< .001	.99	.09 [.07, .12]	.03
Scalar	64.15	10	< .001	.99	.08 [.06, .10]	.03
WF						
Configural	125.86	115	.230	.99	.01 [.00, .02]	.07
Metric	127.14	118	.226	.99	.01 [.00, .02]	.07
Scalar	167.06	117	.002	.98	.02 [.01, .03]	.07
NF						
Configural	147.84	138	.268	.99	.01 [.00, .02]	.07
Metric	169.71	141	.050	.99	.02 [.00, .02]	.07
Partial Metric	156.72	140	.158	.99	.01 [.00, .02]	.07
Scalar	203.76	142	< .001	.98	.02 [.02, .03]	.07
Partial Scalar	156.89	141	.171	.99	.01 [.00, .02]	.07

Note. SIS = Six-Item Screener; B-SEVLT = Brief-Spanish English Verbal Learning Test; WF = Word Fluency; DSS = Digit Symbol Substitution; NF = Neurocognitive Functioning.