Memory and language cognitive data harmonization across the United States and Mexico

Miguel Arce Rentería[1]*, Emily M. Briceño[2]*, Diefei Chen[3,4], Joseph Saenz[5], Lindsay C. Kobayashi[6,7], Christopher Gonzalez[8], Jet M.J. Vonk[9], Richard N. Jones[10], Jennifer J. Manly[1], Rebeca Wong[11], David Weir[6], Kenneth M. Langa[6,12,13], Alden L. Gross[3,4]

[1]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University College of Physicians and Surgeons, New York City, NY, USA, 10032

[2]Department of Physical Medicine & Rehabilitation, University of Michigan Medical School, Ann Arbor, MI, USA, 48108

[3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 2024 E. Monument Street, Baltimore, MD, USA, 21218

[4]Johns Hopkins University Center on Aging and Health, Baltimore, MD, USA, 21218

[5]Edson College of Nursing and Health Innovation at Arizona State University, Phoenix, AZ, USA, 85004

[6]Center for Social Epidemiology and Population Health, Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA, 48108

[7]Survey Research Center, University of Michigan Institute for Social Research, Ann Arbor, MI, USA, 48108

[8]Department of Psychology, Illinois Institute of Technology, Chicago, IL, 60616

[9]Memory and Aging Center, Department of Neurology, University of California San Francisco, San Francisco, CA, USA, 94143

[10]Department of Psychiatry and Human Behavior, Warren Alpert Medical School, Brown University, Providence RI, USA, 02912

[11]Sealy Center on Aging, University of Texas Medical Branch at Galveston, Galveston, TX, USA, 77555

[12]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA, 48108

[13]Veterans Affairs Ann Arbor Center for Clinical Management Research, Ann Arbor, MI,
48108

*Equal contribution, co-first authors and co-corresponding authors

Corresponding Authors:

Miguel Arce Rentería
Department of Neurology
Taub Institute for Research on Alzheimer's Disease and the Aging Brain
Columbia University Medical Center
622 W 168th St
New York, NY, 10032
ma3347@cumc.columbia.edu

Emily M. Briceño
Department of Physical Medicine & Rehabilitation
University of Michigan Medical School
325 E. Eisenhower Blvd
Ann Arbor, MI 48108

Conflict of Interest and Disclosure Statement

None.

Abstract

Introduction: We used cultural neuropsychology-informed procedures to derive and validate harmonized scores representing memory and language across population-based studies in the US and Mexico.

Methods: Data were from the Health and Retirement Study Harmonized Cognitive Assessment Protocol (HRS-HCAP) and the Mexican Health and Aging Study

(MHAS) Ancillary Study on Cognitive Aging (Mex-Cog). We statistically co-calibrated memory and language domains and performed differential item functioning (DIF) analysis using a cultural neuropsychological approach. We examined relationships between harmonized scores, age and education.

Results: We included 3170 participants from the HRS-HCAP [*M*age=76.6 (SD: 7.5), 60% female] and 2042 participants from the Mex-Cog [*M*age=68.1 (SD: 9.0), 59% female]. Five of 7 memory items and 1 of 12 language items demonstrated DIF by study. Harmonized memory and language scores showed expected associations with age and education.

Discussion: A cultural neuropsychological approach to harmonization facilitates the generation of harmonized measures of memory and language function in cross-national studies.

Key Terms: harmonization, cognitive aging, Alzheimer's disease, cross-cultural, cultural neuropsychology

1. Introduction

By 2050, two-thirds of older individuals with dementia will live in low-and middle-income countries (LMICs)[1]. As older adults in LMICs continue to experience longer survival rates and improved healthcare access[2–4] it is critical to understand the factors associated with cognitive decline and dementia risk to address the needs of these aging populations. While the bulk of research on Alzheimer's disease and related dementias (ADRD) comes from high-income countries (i.e., U.S.), cross-national research offers a unique opportunity to understand the sociocultural factors associated with ADRD across individuals residing in the U.S. and LMICs such as Mexico.

The Harmonized Cognitive Assessment Protocol (HCAP) developed through the Health and Retirement Study (HRS) and several of its International Partner

Studies[5] provides a cross-cultural instrument for measuring cognitive function among older adults globally. The HCAP has been implemented in the HRS-HCAP study in the U.S[6] and the Mexican Health and Aging Study (MHAS) Ancillary Study on Cognitive Aging in Mexico (Mex-Cog)[7]. Although the HCAP instruments used in HRS-HCAP and Mex-Cog were designed to optimize comparability, each study has unique methodological and administrative characteristics, and the cohorts have sociocultural, and linguistic differences. These differences required adaptation of HCAP items (i.e., administration and scoring procedures), which complicates the direct comparison of cognitive test scores across studies[8]. Cultural neuropsychological expertise is needed to carefully review these modified neuropsychological instruments to determine whether they are measuring the cognitive construct equivalently across linguistically and culturally diverse populations. Comprehensive data harmonization using a culturally-informed neuropsychological approach is needed for optimal cross-national comparisons of later-life cognitive health using the HCAP.

The HCAP measures several cognitive domains that have been identified with confirmatory factor analysis (CFA), such as memory and language[8–10]. Memory and language abilities are impacted early in the AD process[11,12] and thus are well-suited for the development of harmonized cognitive domain scores.

A critical step in the development of harmonized cognitive scores is their validation. Validation is a complex and multifaceted process needed to ensure that the scores meaningfully represent cognitive health in each study. An initial cross-sectional approach is to examine whether test scores in each study are associated with demographic factors known to be associated with cognitive health in older adults, such as age and educational attainment.

The present study aims to describe the methodology, findings, and an initial validation for harmonized memory and language domain scores across HRS-HCAP and Mex-Cog. We first describe our cultural neuropsychology-informed methodology for the harmonization of these scores. We then examine the measurement equivalence of these scores and perform cross-sectional validation by examining the associations between the harmonized scores with age and education in each study.

2. Methods

2.1 Cohorts

2.1.1 HRS-HCAP. The HRS is an ongoing nationally representative longitudinal study of adults aged 51 years and older living in the U.S[6]. The HRS-HCAP study recruited a randomly selected subsample of adults aged ε65 years who completed the 2016 HRS interview. Details regarding the HRS-HCAP selection process can be found elsewhere[5]. The HRS-HCAP sample includes 2,483 non-Hispanic White participants, 551 non-Hispanic Black participants, 383 Hispanic/Latinx participants, and 79 participants who identified as another race/ethnicity. For the present analysis, we included 3,170 participants, after excluding participants missing the entire HCAP assessment (N=149) and participants who completed the assessment in Spanish (n=177; given the small sample size that would impact the reliability of DIF analyses and to provide a more controlled comparison to the Mex-Cog sample).

2.1.2 Mex-Cog. MHAS is a nationally representative sample of adults 50 years of age and older living in Mexico[7]. Mex-Cog participants were a subsample of adults aged ε55 years who completed the 2015 MHAS wave. Mex-Cog study selection procedures are available elsewhere[7,13]. In brief, stratified sampling procedures were used to select a subsample of MHAS participants from eight

Mexican states using criteria evaluating the distribution of socioeconomic factors (percent urban/rural, history of return migration from the U.S.) and health characteristics (obesity, diabetes, mine industry, pottery industry). Mex-Cog includes 2,265 participants, of which 2,042 were administered the HCAP.

2.2 Cognitive Assessment

The HCAP battery was designed to assess the cognitive domains of memory, language, orientation, visuospatial, and executive functioning. Details regarding the cognitive tests included in HCAPs have been published previously[5,7,8]. For the current study, we included items measuring the domains of memory and language as determined by the cognitive factor structure of the HCAP[9,10,14]. Table 1 lists all test included in our memory and language domains.

2.3 Procedures

2.3.1 Harmonization of demographic variables. Age, years of schooling, and sex/gender were collected via self-report for both studies. Education was further harmonized according to the 2011 International Standard Classification of Education (ISCED) for the purposes of sample characterization and adjustment in DIF analyses[15].

2.3.2 Pre-statistical harmonization. To determine candidate linking items between the two studies, we applied a cultural neuropsychological approach to pre-statistical harmonization of cognitive data given the cultural and linguistic differences between the two cohorts; these procedural details are available elsewhere[8]. Briefly, neuropsychologists (EMB, MAR) collaboratively reviewed all memory and language items for cross-study comparability in conjunction with study team members with competence in the languages and cultures represented in the two cohorts. Comparability was evaluated across 1) administration and scoring procedures, 2) coding procedures, and 3) linguistic and cultural equivalence. Linguistic equivalence was determined by evaluating the translation of test instructions and items (e.g., translated words are of similar linguistic frequency, translated instructions are of comparable clarity and complexity.). For cultural equivalence we considered the degree of similar cultural familiarity of the items and the construct equivalence of the item from a theoretical perspective. For instance, we considered the degree to which the Spanish version of the story recall included details that were as culturally familiar as the English version. After review for comparability, potential linking items were classified as either “confident” (i.e., no known features violating item comparability) or as “tentative” (i.e., possible features that may violate item comparability) linking items. As an example of both linguistic and cultural equivalence considerations, we noted that the sentence repetition item was linguistically slightly more challenging in English (i.e., sentence includes several plural words such that if the “s” is not pronounced, the item is scored as incorrect) and it represented a culturally more common phrase in Mexico compared to the phrase used in the U.S., potentially making the item easier in Spanish. Thus, we classified it as a “tentative” linking item. Items determined to be non-comparable across cohorts were treated as unique items. An example of a unique item was the CERAD Word List Memory Test[16], given that in HRS-HCAP the test stimuli were presented visually and the list of words alternated order for each trial, whereas in Mex-Cog the words were presented verbally and in the same order each trial.

2.3.3 Statistical harmonization. We conducted statistical harmonization using an item banking approach[17]. Separately for each cognitive domains, we estimated a CFA model through ML in the HRS-HCAP using all available test items for the domain. For model identification, the mean and variance of the latent variable were fixed to 0 and 1, respectively. Two parameters (factor loadings, thresholds/intercepts) were estimated from this model for each cognitive test item and were saved into an item bank. Factor loading describes the strength of association between the item and the underlying trait (memory or language). Thresholds or intercepts reflect the average level of the underlying trait at which the item is most discriminating.

After estimating memory and language CFAs in HRS-HCAP, we estimated similar CFAs in Mex-Cog. In this model, we standardized the latent variable to be on the scale of HRS-HCAP by leveraging parameters saved from the first round of estimation (i.e., item factor loadings and thresholds). Parameters for items in Mex-Cog seen in HRS-HCAP were fixed to their values in HRS-HCAP, while the mean and variance of the latent variable in the Mex-Cog model were freely estimated, as well as the item parameters for Mex-Cog items not yet in the item bank[17].

In a final score-generating model for each domain, we pooled all participants to estimate one CFA model for that domain, in which we placed constraints on all item parameters corresponding to their previously estimated values. From these models, we estimated the non-DIF-adjusted factor scores representing memory and language.

Model fit was considered perfect if CFI=1 and RMSEA=0 and SRMR=0, good if CFI≥0.95 and RMSEA ≤0.05 and SRMR≤0.05, adequate if CFI≥0.90 and RMSEA≤0.08 and SRMR≤0.08, and poor if either CFI<0.9 or RMSEA>0.08 or SRMR>0.08.

2.4 Differential item functioning (DIF)

To empirically test the assumptions of equivalence of common items from the harmonization procedure, we tested for DIF by study using item response theory methods[18]. Modeled on a previously published study[8], we used a Multiple-Indicator, Multiple-Cause (MIMIC) model[19]. In this study, MIMIC models were adjusted for age, sex, and education. This approach estimates CFA models with categorical response variables (i.e., cognitive test items) as factor indicators and a grouping variable for study membership (HRS-HCAP vs. Mex-Cog) as a predictor of the latent response variable. Starting with a baseline CFA model without modeling the direct effect of group membership on the latent response variables, stepwise forward selection leverages model modification indices to select direct effects of the grouping variable on an item to be added to the model. Direct paths are added between study membership and items, until no statistically significant modification indices remain (defined at $p<0.05$).

Separately for each domain, we first used the MIMIC model approach to test for non-negligible DIF only among confident linking items. We defined non-negligible DIF to be present if a DIF effect estimate falls outside a pre-defined caliper for small effects (i.e., the 95% confidence interval of the odds ratio for the direct effect is between 0.66 and 1.5 in a multivariate probit regression model)[20]. While some confident linking items might show DIF, most are not expected to have DIF because experts decided they were unlikely to show any based on cultural and linguistic features. When evaluating for DIF among confident linking items in the language domain, we constrained one item (animal naming) to be free from DIF across studies. We did so because it was the only continuously distributed item for the language domain. Next, using confident linking items that exhibited no or negligible DIF as anchors to link the studies, we conducted DIF detection among the tentative linking items[10]. The same DIF detection procedures were repeated separately for the memory and language domains.

For items exhibiting uniform DIF (a difference in thresholds or intercepts), we computed an odds ratio for the strength of the association between cohort (reference: HRS-HCAP) and the item. For items exhibiting non-uniform DIF (difference in item factor loading), we computed the difference in loadings between the Mex-Cog and the HRS-HCAP.

2.5 Evaluation of salient DIF

DIF detection may yield evidence for *statistically significant* DIF, which may or may not be *impactful* on the resulting domain-specific scores. After the DIF detection procedure, we estimated DIF-adjusted factor scores by allowing the items identified with DIF to have different item parameters across studies. We evaluated *salient* DIF by comparing the distribution of the DIF-adjusted scores with non-DIF-adjusted scores. We calculated the proportion of participants whose DIF-adjusted scores differed from non-DIF adjusted scores by more than 0.3 SD units[21,22]. Finally, we evaluated test information curves from these final DIF-adjusted models between HRS-HCAP and Mex-Cog.

2.6 Validation of harmonized factor scores

For criterion validation, we evaluated how age and educational attainment were related to the harmonized factor scores by study. Correlation coefficients were calculated to indicate associations with continuous age and years of schooling. Means and z-scores were calculated to indicate association with the categorical education variable.

Descriptive analyses and data management were conducted in Stata version 17[23]. IRT and MIMIC modeling were conducted using Mplus version 8.2[24].

3. Results

Table 2 describes the sociodemographic characteristics of participants in each study. On average, the HRS-HCAP participants were older, and with higher educational attainment compared to the Mex-Cog participants. There were no differences in sex/gender between the two studies.

3.1 Pre-statistical harmonization*.* Table 1 shows the items included in the memory and language domains. Using our cultural neuropsychological approach, we identified 1 confident and 6 tentative anchor items in the memory domain, and 6 confident and 6 tentative anchor items in the language domain. The remaining 3 items for memory and 4 items for language were determined to be unique items within each study.

3.2 Domain Score Model Fit. Table 3 displays the factor loadings and item thresholds/intercepts for each item in the memory and language domains. Absolute model fit in HRS-HCAP was excellent for memory (RMSEA=0.044; CFI=0.981; SRMR=0.023) and for language (RMSEA=0.020; CFI=0.971; SRMR=0.071). Absolute model fit in Mex-Cog, without model constraints, was also excellent for memory (RMSEA=0.048; CFI=0.985; SRMR=0.033) and good for language (RMSEA=0.026; CFI=0.964; SRMR=0.085).

3.3 Memory Domain

3.3.1 DIF Results. Table 4 displays DIF results for the cross-study linking items. Given that only one item in the memory domain was a confident linking item (CERAD constructional praxis delayed recall), we could not evaluate DIF for this item and thus were required to constrain it as a cross-study anchor (i.e., constrain the item to not show DIF). Five of the six tentative linking items for the memory domain exhibited non-negligible DIF. Four of these items showed uniform DIF and one showed non-uniform DIF. One item (3-word delayed recall) demonstrated negligible

DIF, indicating that it measured the memory domain in a similar fashion in HRS-HCAP and Mex-Cog.

3.3.2 Salient DIF. Among the Mex-Cog participants, 5.7% (n=116) of the sample had non-DIF adjusted scores that were 0.3 SD units greater than their DIF-adjusted scores (Figure 1). These results indicate that not accounting for DIF would lead to underestimation of memory scores for 5.7% of participants in the Mex-Cog study.

3.3.3 Measurement Precision. Information curves for memory domain scores showed excellent reliability (*r*>0.90) of the memory domain across most of the distribution of the latent trait for the HRS-HCAP (-3.3<z< 2.5; Figure 2). For the Mex-Cog, reliability was lower (*r*<0.90) at the low end (z<-1.9) of the latent trait, and excellent (*r*>0.90) for the higher end of the latent ability level (1.9<z<3.2).

3.4 Language Domain

3.4.1 DIF results. After constraining the Animal naming item to be free from DIF across studies, we observed non-negligible DIF in one of the five confident linking items (elbow naming; non-uniform DIF), and one of six tentative linking items hammer naming; uniform DIF).

3.4.2 Salient DIF. DIF adjustment had a minimal effect on scores, such that N=53 (2.6%) of Mex-Cog participants had DIF-adjusted scores that differed by more than 0.3 SD from the non-DIF-adjusted scores (Figure 1).

3.4.3 Measurement precision. Information curves for the language domain showed that both studies exhibited relatively better reliability at the low end of the latent trait, whereas reliability was low (*r*<0.8) at higher levels of the latent trait (*z*>-0.9 for HRS-HCAP and *z*>-1.5 for Mex-Cog; Figure 2). This low reliability occurred at latent trait levels that were the most common in both studies (90.0% of HRS-HCAP and 93.1% of Mex-Cog sample).

3.5 Age, education, and harmonized scores

As expected, there were negative associations with age and positive associations with education for both memory and language domains in both studies (Figure 3).

5. Discussion

By applying a cultural neuropsychological approach to cross-national cognitive data harmonization, we developed memory and language domain factor scores for cross-national comparisons of cognitive functioning between HCAP studies in the United States (HRS-HCAP) and Mexico (Mex-Cog). We observed measurement differences in the harmonized memory and language scores that impacted few participants in the HRS-HCAP and Mex-Cog, suggesting that cognitive performance is measured comparably in each study by the HCAP. Memory demonstrated strong measurement precision across all levels of the latent ability for both studies. However, the language domain demonstrated lower measurement precision, particularly at higher levels of the latent trait. Lastly, initial validation of these harmonized scores demonstrated similar and expected associations with age and education across studies.

There have been previous efforts that leverage ongoing international longitudinal studies for cross-national studies of cognitive aging and ADRD[5,24,25]. These studies have linked several sociodemographic and health factors with increased risk of cognitive impairment and decline across several countries in various continents[26–30]. Our study builds upon this prior work in various ways. First, efforts such as the COSMIC consortium have relied on standardizing scores across studies[29,30]. Standardization of scores may bias harmonized analyses because it eliminates possible differences in distributions of scores between studies[31]. Additionally, standardized scores do not account for differences in measurement precision between instruments across studies, or within instruments across cultural and linguistic groups[32]. By combining a cultural neuropsychological approach to pre-statistical harmonization, advanced structural equation modeling in statistical harmonization, and evaluation of differential item functioning, we can more reliably equate scores across international cohorts[8].

Other cross-national cognitive data harmonization efforts have come from the HRS and its International Partner Studies[5]. These studies have relied on briefer measures of global cognitive functioning rather than a comprehensive neuropsychological assessment. Prior work has provided support for the cognitive factor structure of the HCAP battery[9,10,14], which allowed us to characterize memory and language with various items for each domain. Memory and language abilities are particularly relevant for cognitive aging and may be more sensitive to subtle cognitive decline than measures of global cognition[11,33,34]. As such, our study provides a foundation from which to evaluate two of the earlier cognitive markers of AD across two economically and culturally distinct countries. Furthermore, composite domain scores provide a more robust measurement of cognition and improve our ability to detect change over time compared to individual test scores[35].

Evaluation of measurement precision of the memory and language domain scores revealed strengths and weaknesses that informs their use in future studies. While a strength of the HCAP is that its use allows the HRS-HCAP and Mex-Cog to administer largely the same comprehensive battery, necessary modifications were made to improve the linguistic, educational, and cultural appropriateness of the

HCAP for the Mex-Cog among Spanish-speaking older adults and those with little to no formal schooling[8]. Despite these modifications, the harmonized memory domain scores had reduced measurement precision in Mex-Cog at lower levels of ability, whereas measurement precision was high across the range of ability for HRS-HCAP. Although we adjusted for education in our DIF analyses, education may still impact measurement, particularly at the low end of the ability range in Mexico. The Mex-Cog sample has a higher prevalence of people with limited schooling (51% with none or without primary education) than the HRS-HCAP sample (0.3%). As such, while we adjusted for years of schooling, there may be differences in educational quality and level of literacy impacting the reliability of the memory score at lower levels of the latent trait in Mex-Cog. Future waves of the Mex-Cog study may consider incorporating additional memory items, such as a recognition task, that may be more sensitive to the lower end of ability, as well as measures to characterize quality of education in both the HRS and Mex-Cog.

The language domain, in contrast, showed relatively better measurement precision at the lower end of ability in both cohorts. This result was expected, as the language items largely consisted of simple naming and comprehension items designed to capture significant aphasia[36]. As a result, the language domain scores are useful for measuring the very low/impaired end of the language ability range (e.g., aphasia) but they have limited utility in measuring language ability in the absence of clinical impairment. Future applications of these scores may be more appropriate for developing impairment cut scores for diagnostic classification (mild cognitive impairment, dementia) rather than as a continuous variable. Future iterations of the HCAP battery may consider expanding the language items to better capture the upper end of the ability range, such as including more challenging confrontation naming tests and/or additional measures of verbal fluency[37,38].

The present study had several strengths. We utilized data from two well-characterized cohorts of older adults in the U.S. and Mexico. We used a multidisciplinary, cultural neuropsychological approach to pre-statistical harmonization to minimize bias in statistical harmonization of memory and language domain scores[8]. Our statistical harmonization process capitalized on both common and unique items across studies to maximize measurement precision[8,19]. We examined the degree of measurement equivalence of our domain scores across studies using DIF, adjusting for age and educational attainment.

In terms of limitations, although we carefully selected our cross-study linking items using all available information, we cannot rule out the possibility of undocumented item differences across studies. We accounted for this possibility by evaluating for measurement equivalence using DIF analyses in all items except for two. We were unable to evaluate for DIF in the CERAD constructional praxis delay because it was the *only* memory linking item classified as "confident" and thus were required to treat it as a cross-study anchor item in the DIF analyses. Similarly, semantic fluency was ineligible for DIF analysis because of its large variance compared to other dichotomous items in the domain. Prior studies have found measurement equivalence for semantic fluency when administered between English and Spanish speakers[39], thus reducing the concern of unaccounted DIF biasing our language factor scores. Regardless, additional work is needed to evaluate the assumption of measurement equivalence of these anchor items.

Our study provides a strong foundation for future cross-national investigations of cognitive aging and ADRD. Future studies can utilize these harmonized cognitive scores to investigate determinants of late-life cognitive decline and dementia in the U.S. and Mexico. Given the cultural and linguistic differences across cross-national cohorts and their relevance to cognitive measurement, a cultural neuropsychological approach is necessary for reliable and valid inferences about cognitive health across national contexts. Continued cross-national investigation of the factors that increase risk and/or resilience to dementia will aid in understanding how to mitigate the impact of dementia globally.

Consent Statement:

Informed consent was obtained from all HRS-HCAP participants and/or their surrogates, and study protocols were approved by the University of Michigan Institutional Review Board (IRB). All Mex-Cog participants provided informed consent and study protocols were approved by the IRBs of the University of Texas Medical Branch (U.S.) and the National Institute of Public Health and the National Institute of Statistics and Geography (Mexico).

Acknowledgments:

This research was completed with funding from the NIH/HHS (5-R24-AG-065182-02, EB,MAR, KL, DW), and NIA (K99AG066932, MAR; P30AG024824, EB; R01AG070953, LB,AG; R01AG030153; JS R00AG058799). The HRS and HCAP are funded by the National Institute on Aging (U01 AG009740 and U01 AG058499, respectively) and performed at the Institute for Social Research, University of Michigan. HRS-HCAP data and documentation are available at https://hrs.isr.umich.edu/data-products/hcap. The MHAS is funded by the NIA/NIH (grant R01 AG018016) and the Instituto Nacional de Estadística y Geografia (INEGI)

in Mexico. The Mex-Cog is sponsored by the NIA/NIH (R01 AG051158). Data files and documentation are public use and available at www.MHASwewb.org

Funding:

NIH/NIA: 5-R24-AG-065182-02, K99AG066932, P30AG024824, R01AG070953, R01AG030153, R00AG058799, U01 AG009740, U01 AG058499, R01 AG018016, R01 AG051158.

References

1. Patterson C. *World Alzheimer Report 2018. The State of the Art of Dementia Research: New Frontiers. An Analysis of Prevalence, Incidence, Cost and Trends.*; 2018.
2. Bendavid E, Bhattacharya J. The Relationship of Health Aid to Population Health Improvements. *JAMA Intern Med*. 2014;174(6):881. doi:10.1001/jamainternmed.2014.292
3. Miladinov G. The mechanism between mortality, population growth and ageing of the population in the European lower and upper middle income countries. *PLoS One*. 2021;16(10):e0259169. doi:10.1371/journal.pone.0259169
4. Sudharsanan N, Bloom DE, Sudharsanan N. The demography of aging in low- and middle-income countries: Chronological versus functional perspectives. In: *In Future Directions for the Demography of Aging: Proceedings of a Workshop*. ; 2018:309-338.
5. Langa KM, Ryan LH, McCammon RJ, et al. The Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: Study Design and Methods. *Neuroepidemiology*. 2020;54(1):64-74. doi:10.1159/000503004
6. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol*. 2014;43(2):576-585. doi:10.1093/ije/dyu067
7. Wong R, Michaels-Obregon A, Palloni A. Cohort Profile: The Mexican Health and Aging Study (MHAS). *Int J Epidemiol*. 2017;46(2):e2-e2. doi:10.1093/ije/dyu263
8. Briceño EM, Arce Rentería M, Gross AL, et al. A cultural neuropsychological approach to harmonization of cognitive data across culturally and linguistically diverse older adult populations. *Neuropsychology*. 2023;37(3):247-257. doi:10.1037/neu0000816
9. Arce Rentería M, Manly JJ, Vonk JMJ, et al. Midlife Vascular Factors and Prevalence of Mild Cognitive Impairment in Late-Life in Mexico. *Journal of the International Neuropsychological Society*. 2022;28(4):351-361. doi:10.1017/S1355617721000539

10. Jones R, Manly JJ, Ryan L, Levine D, McCammon R, Weir D. Factor structure of the Harmonized Cognitive Assessment Protocol neuropsychological battery in the Health and Retirement Study. *Journal of the International Neuropsychological Society*. In Press.
11. Weintraub S, Carrillo MC, Farias ST, et al. Measuring cognition and function in the preclinical stage of Alzheimer's disease. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. 2018;4(1):64-75. doi:10.1016/j.trci.2018.01.003
12. Dubois B, Hampel H, Feldman HH, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*. 2016;12(3):292-323. doi:10.1016/j.jalz.2016.02.002
13. Mejia-Arango S, Nevarez R, Michaels-Obregon A, et al. The Mexican Cognitive Aging Ancillary Study (Mex-Cog): Study Design and Methods. *Arch Gerontol Geriatr*. 2020;91:104210. doi:10.1016/j.archger.2020.104210
14. Gross AL, Khobragade PY, Meijer E, Saxton JA. Measurement and Structure of Cognition in the Longitudinal Aging Study in India–Diagnostic Assessment of Dementia. *J Am Geriatr Soc*. 2020;68(S3). doi:10.1111/jgs.16738
15. UNESCO Institute for Statistics. *International Standard Classification of Education (ISCED) 2011*. UNESCO Institute for Statistics; 2012. doi:10.15220/978-92-9189-123-8-en
16. Morris JC, Heyman A, Mohs RC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assesment of Alzheimer's disease. *Neurology*. 1989;39(9):1159-1159. doi:10.1212/WNL.39.9.1159
17. Vonk JMJ, Gross AL, Zammit AR, et al. Cross-national harmonization of cognitive measures across HRS HCAP (USA) and LASI-DAD (India). *PLoS One*. 2022;17(2):e0264166. doi:10.1371/journal.pone.0264166
18. Camilli G, Shepard LA, Shepard L. *Methods for Identifying Biased Test Items* . Vol 4. Sage Publications; 1994.
19. Jones RN. Identification of Measurement Differences Between English and Spanish Language Versions of the Mini-Mental State Examination. *Med Care*. 2006;44(Suppl 3):S124-S133. doi:10.1097/01.mlr.0000245250.50114.0f
20. Zwick R. A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*. 2012;2012(1):i-30. doi:10.1002/j.2333-8504.2012.tb02290.x
21. Crane PK, Gibbons LE, Narasimhalu K, Lai JS, Cella D. Rapid detection of differential item functioning in assessments of health-related quality of life: The Functional Assessment of Cancer Therapy. *Quality of Life Research*. 2007;16(1):101-114. doi:10.1007/s11136-006-0035-7
22. Goel A, Gross A. Differential item functioning in the cognitive screener used in the Longitudinal Aging Study in India. *Int Psychogeriatr*. 2019;31(9):1331-1341. doi:10.1017/S1041610218001746
23. StataCorp. Stata Statistical Software: Release 17. Published online 2021.
24. Muthén LK, Muthén BO. Mplus: Statistical Analysis with Latent Variables: User's Guide. Published online 2017.

25. Briceño EM, Gross AL, Giordani BJ, et al. Pre-Statistical Considerations for Harmonization of Cognitive Instruments: Harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS. *Journal of Alzheimer's Disease*. 2021;83(4):1803-1813. doi:10.3233/JAD-210459
26. Crimmins EM, Kim JK, Langa KM, Weir DR. Assessment of Cognition Using Surveys and Neuropsychological Assessment: The Health and Retirement Study and the Aging, Demographics, and Memory Study. *J Gerontol B Psychol Sci Soc Sci*. 2011;66B(Supplement 1):i162-i171. doi:10.1093/geronb/gbr048
27. Downer B, Veeranki SP, Wong R. A Late Life Risk Index for Severe Cognitive Impairment in Mexico. *Journal of Alzheimer's Disease*. 2016;52(1):191-203. doi:10.3233/JAD-150702
28. McEvoy CT, Guyer H, Langa KM, Yaffe K. Neuroprotective Diets Are Associated with Better Cognitive Function: The Health and Retirement Study. *J Am Geriatr Soc*. 2017;65(8):1857-1862. doi:10.1111/jgs.14922
29. Röhr S, Pabst A, Riedel-Heller SG, et al. Estimating prevalence of subjective cognitive decline in and across international cohort studies of aging: a COSMIC study. *Alzheimers Res Ther*. 2020;12(1):167. doi:10.1186/s13195-020-00734-y
30. Sachdev PS, Lipnicki DM, Kochan NA, et al. The Prevalence of Mild Cognitive Impairment in Diverse Geographical and Ethnocultural Regions: The COSMIC Collaboration. *PLoS One*. 2015;10(11):e0142388. doi:10.1371/journal.pone.0142388
31. Griffith LE, van den Heuvel E, Raina P, et al. Comparison of Standardization Methods for the Harmonization of Phenotype Data: An Application to Cognitive Measures. *Am J Epidemiol*. 2016;184(10):770-778. doi:10.1093/aje/kww098
32. Ramirez M, Ford ME, Stewart AL, A. Teresi J. Measurement Issues in Health Disparities Research. *Health Serv Res*. 2005;40(5p2):1640-1657. doi:10.1111/j.1475-6773.2005.00450.x
33. Snowden JS, Stopford CL, Julien CL, et al. Cognitive Phenotypes in Alzheimer's Disease and Genetic Risk. *Cortex*. 2007;43(7):835-845. doi:10.1016/S0010-9452(08)70683-X
34. Taler V, Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *J Clin Exp Neuropsychol*. 2008;30(5):501-556. doi:10.1080/13803390701550128
35. Jonaitis EM, Koscik RL, Clark LR, et al. Measuring longitudinal cognition: Individual tests versus composites. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*. 2019;11(1):74-84. doi:10.1016/j.dadm.2018.11.006
36. Goodglas H. *The Assessment of Aphasia and Related Disorders*. Vol 2. Lea & Febiger; 1983.
37. Ma Y, Carlsson CM, Wahoske ML, et al. Latent Factor Structure and Measurement Invariance of the NIH Toolbox Cognition Battery in an Alzheimer's Disease Research Sample. *Journal of the International Neuropsychological Society*. 2021;27(5):412-425. doi:10.1017/S1355617720000922
38. Avila JF, Rentería MA, Witkiewitz K, Verney SP, Vonk JMJ, Manly JJ. Measurement invariance of neuropsychological measures of cognitive aging across race/ethnicity by sex/gender groups. *Neuropsychology*. 2020;34(1):3-14. doi:10.1037/neu0000584

39. Siedlecki KL, Manly JJ, Brickman AM, Schupf N, Tang MX, Stern Y. Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology*. 2010;24(3):402-411. doi:10.1037/a0017515

Table 1. Items included in Memory and Language Domains

Domain	Item	Study	Variable Type	Linking item Confidence
Memory				
	CERAD Constructional praxis delay	HRS-HCAP Mex-Cog	Continuous	Confident
	CERAD Word List Immediate sum of 3 trials	HRS-HCAP Mex-Cog	Continuous	No link
	CERAD Word List Delay	HRS-HCAP Mex-Cog	Continuous	No Link
	CERAD Word List Recognition	HRS-HCAP Mex-Cog	Continuous	No Link
	WMS-IV Logical Memory Immediate Recall	HRS-HCAP Mex-Cog	Continuous	Tentative
	WMS-IV Logical Memory Delayed Recall	HRS-HCAP Mex-Cog	Continuous	Tentative
	WMS-IV Logical Memory Recognition	HRS-HCAP	Continuous	No link
	East Boston Memory Test (Brave man) Immediate Recall	HRS-HCAP Mex-Cog	Categorical	Tentative
	East Boston Memory Test (Brave man) Delayed Recall	HRS-HCAP Mex-Cog	Categorical	Tentative
	3 Word Immediate Recall	HRS-HCAP Mex-Cog	Categorical	Tentative
	3 Word Delayed Recall	HRS-HCAP Mex-Cog	Categorical	Tentative
Language/Fluency				
	TICS – Naming (cactus)	HRS-HCAP	Categorical	No link
	TICS – Naming (scissors)	HRS-HCAP	Categorical	Tentative

	Mex-Cog		
Naming (common object)	HRS-HCAP Mex-Cog	Categorical	Tentative
Naming (writing utensil)	HRS-HCAP Mex-Cog	Categorical	Confident
1066 - Naming (elbow)	HRS-HCAP Mex-Cog	Categorical	Confident
Read and following command	HRS-HCAP Mex-Cog	Categorical	Confident
Follow command (R does not read)*	HRS-HCAP	Categorical	No link
1066 – Following instructions	HRS-HCAP Mex-Cog	Categorical	Confident
Following instructions – 3 steps	HRS-HCAP Mex-Cog	Categorical	Confident
TICS – Name current president	HRS-HCAP	Categorical	No link
Animal Fluency	HRS-HCAP Mex-Cog	Continuous	Confident
Write a sentence	HRS-HCAP Mex-Cog	Categorical	Tentative
Repetition of phrase	HRS-HCAP Mex-Cog	Categorical	Tentative
1066 – What does one do with a hammer	HRS-HCAP Mex-Cog	Categorical	Tentative
1066 – Where is the local market?	HRS-HCAP Mex-Cog	Categorical	Tentative
Definition (Bridge)	Mex-Cog	Categorical	No Link

Note. CERAD = Consortium to Establish a Registry for Alzheimer's Disease; HRS-HCAP is Health and Retirement Study Harmonized Cognitive Assessment Protocol. TICS = Telephone Interview for Cognitive Status. "Categorical" refers to both ordinal and binary variables. *Refer to Briceño & Arce Rentería et al. (2022) for additional details regarding pre-statistical harmonization of this item.

Table 2. Sociodemographic and Health Characteristics of Participants in HRS-HCAP and Mex-Cog

Characteristic	HRS-HCAP (*n* = 3170)	Mex-Cog (*n* = 2042)	*p*
	Mean (SD) or N (%)	Mean (SD) or N (%)	
Age, mean (SD)	76.7 (7.5)	68.1 (9.0)	<0.001
Female, n (%)	1919 (60.5)	1203 (58.9)	0.243
Education, n (%)			<0.001
None or Early Childhood Education	8 (0.3)	1023 (50.5)	
Primary education (US grades 1-6)	68 (2.2)	452 (22.3)	
Lower secondary education (US grades 7-9)	419 (13.2)	317 (15.7)	
Upper secondary education (US grades 10-12)	1725 (54.5)	60 (3.0)	
Any college	948 (29.9)	172 (8.5)	

Note. HRS-HCAP is Health and Retirement Study Harmonized Cognitive Assessment Protocol.

Table 3. Factor loadings and thresholds or intercepts for Memory and Language from the CFA models

Indicators	**Factor loading**		**Threshold or intercept**		**Data source**
	Raw	**Standardized**	**Threshold #**		
Memory					
CERAD Word List Immediate sum	4.44	0.84		17.54	HRS-HCAP only
CERAD Word List Immediate sum	5.28	0.83		17.45	Mex-Cog only
WMS-IV Logical Memory Immediate	3.41	0.67		9.94	Both
WMS-IV Logical Memory Delayed Recall	3.62	0.67		7.52	Both
3-Word Delayed Recall	0.73	0.73	1	-1.85	Both
			2	-1.31	
			3	-0.46	
CERAD Word List Delayed	2.28	0.86		5.18	HRS-HCAP

Recall					only
CERAD Word List Delayed Recall	2.58	0.80		5.54	Mex-Cog only
CERAD Constructional Praxis Delay	2.13	0.66		5.89	Both
CERAD Word List Recognition	1.72	0.70		18.56	HRS-HCAP only
CERAD Word List Recognition	2.84	0.65		19.20	Mex-Cog only
WMS-IV Logical Memory Recognition	1.56	0.57		10.38	HRS-HCAP only
East Boston Memory Test Delayed	0.57	0.57	1	-0.58	Both
			2	-0.09	
			3	0.43	
			4	1.03	
			5	1.83	
			6	2.50	
East Boston Memory Test Immediate	0.45	0.45	1	-1.66	Both
			2	-0.90	
			3	-0.16	
			4	0.52	
			5	1.23	
			6	2.02	
3-Word Immediate Recall	0.47	0.47	1	-2.08	Both
			2	-1.33	
Language					
Animal fluency	4.64	0.70		16.05	Both
TICS – Naming (Cactus)	0.80	0.80	1	-1.42	HRS-HCAP only
TICS – Naming (Scissors)	0.74	0.74	1	-2.11	Both
TICS – Naming (Watch)	0.78	0.78	1	-2.57	Both
Naming (Writing Utensil)	0.67	0.67	1	-2.46	Both
1066 – Naming (Elbow)	0.86	0.86	1	-2.25	Both
Write a Sentence	0.61	0.61	1	-1.55	Both
Read and Follow Command	0.61	0.61	1	-1.96	Both
Repetition of phrase	0.46	0.46	1	-0.50	Both
1066 – What Does One Do with a Hammer	0.40	0.40	1	-1.42	Both
Definition (Bridge)	0.66	0.52	1	-1.17	Mex-Cog only
1066 – Following Instructions	0.85	0.85	1	-2.32	Both
1066 – Where is the Local Market?	0.55	0.55	1	-0.88	Both
Following Instructions 3 Step	0.37	0.37	1	-2.59	Both
			2	-1.90	

			3	-0.63	
TICS - Name Current President	0.84	0.84	1	-1.61	HRS HCAP only

Note: Raw factor loadings, thresholds/intercepts are reported, standardized parameters are included in the parentheses. CERAD = Consortium to Establish a Registry for Alzheimer's Disease; HRS-HCAP is Health and Retirement Study Harmonized Cognitive Assessment Protocol. TICS = Telephone Interview for Cognitive Status.

Table 4. Differential Item Functioning (DIF) Results across Memory and Language Domains

Cognitive Domain	**Stage of DIF testing**	**Test Item**	**Variable Type**	**Type of DIF identified via MIMIC**	**Uniform DIF: Association with cohort (REF: HRS-HCAP)**		**Non-uniform DIF: Difference in loading (Mex-Cog & HRS-HCAP)**	
					Odds Ratio	95% CI	Difference	95% CI
Memory								
	DIF Among Confident Items							
		CERAD Constructio Praxis Dela	Continuous	N/A				
	DIF Among Tentative Items, Treating Confident Items as Anchors							
		WMS-IV Logical Memory Immediate Recall	Continuous	Uniform	-2.716	(-2.941, -2.491)		
		WMS-IV Logical Memory Delayed Recall	Continuous	Uniform	-1.727	(-1.980, -1.474)		
		3-Word Delayed Recall	Categorical	Uniform	0.803	(0.753, 0.857)		
		East Boston Memory Test (Brave	Categorical	Uniform	1.692	(1.596, 1.795)		

		man) Delayed Recall				
		East Boston Memory Test (Brave man) Immediate Recall	Categorical	Non-uniform	0.444	(0.35, 0.538)
		3-Word Immediate Recall	Categorical	Negligible		
Language						
	DIF among confident items					
		Animal fluency	Continuous	N/A		
		Naming (Writing Utensil)	Categorical	Negligible		
		1066 – Naming (Elbow)	Categorical	Non-uniform	-0.297	(-0.489, -0.105)
		Read and Follow Command	Categorical	Negligible		
		1066 – Following Instructions		Negligible		
		Following Instructions 3 Step	Categorical	Negligible		
	DIF among tentative items, treating confident items as anchors					
		TICS – Naming (Scissors)	Categorical	Negligible		
		Naming (common object)	Categorical	Negligible		
		Write a	Categoric	Negligibl		

Sentence	al	e		
Repetition of phrase	Categorical	Negligible		
1066 – What Does One Do with a Hammer	Categorical	Uniform	2.083	(1.834, 2.366)
1066 – Where is the Local Market?	Categorical	Negligible		

Note. Reference group is HRS-HCAP. The Odds Ratio (OR) is the difference (on an odds scale) in outcome between and Mex-Cog and HRS-HCAP, adjusting for the latent ability. The Odds Ratio (OR) is the difference (on an odds scale) in outcome between and Mex-Cog and HRS-HCAP, adjusting for the latent ability. Coefficients greater than 1 (for the OR) or 0 (for the difference) implies better performance than expected on the item in Mex-Cog, compared to HRS-HCAP, whereas a coefficient less than 1 (for the OR) or 0 (for the difference) indicates better performance on the item than expected in HRS-HCAP, compared to Mex-Cog. DIF among tentative items, treating confident items as anchors. CERAD = Consortium to Establish a Registry for Alzheimer's Disease; HRS-HCAP is Health and Retirement Study Harmonized Cognitive Assessment Protocol; TICS = Telephone Interview for Cognitive Status

Figure 1. The curves represent the distributions of standardized differences between non-DIF-adjusted and DIF-adjusted scores by study, for the domains of memory (Panel A) and language (Panel B), respectively.

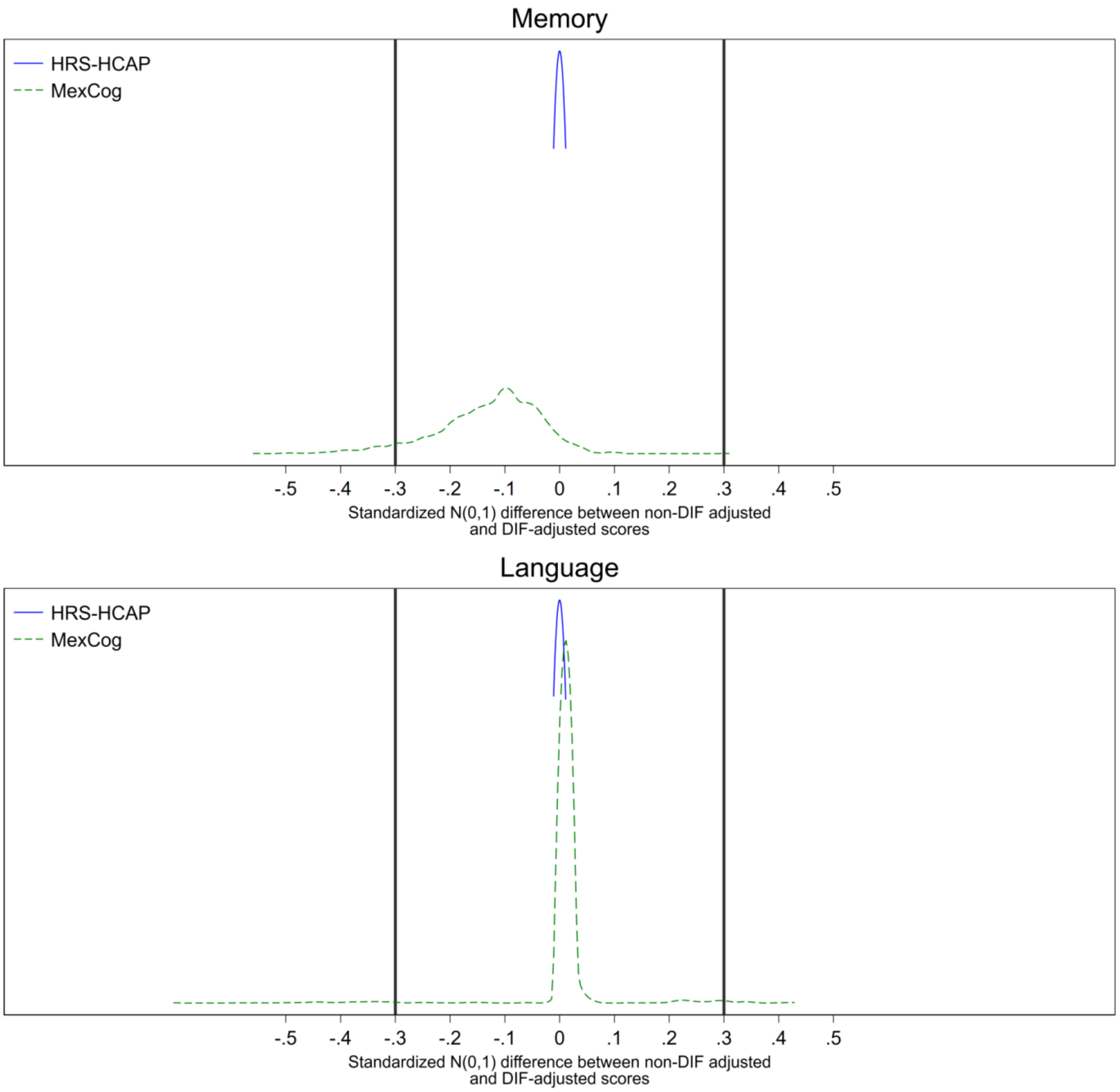


Figure 2. The top halves of the plots represent the reliability of factor scores, and the bottom halves are histograms of factor scores for memory (Panel A) and language (Panel B) domains by study. The goal of this figure is to illustrate the change in the reliability of estimated factor scores as a function of corresponding levels on the latent trait.

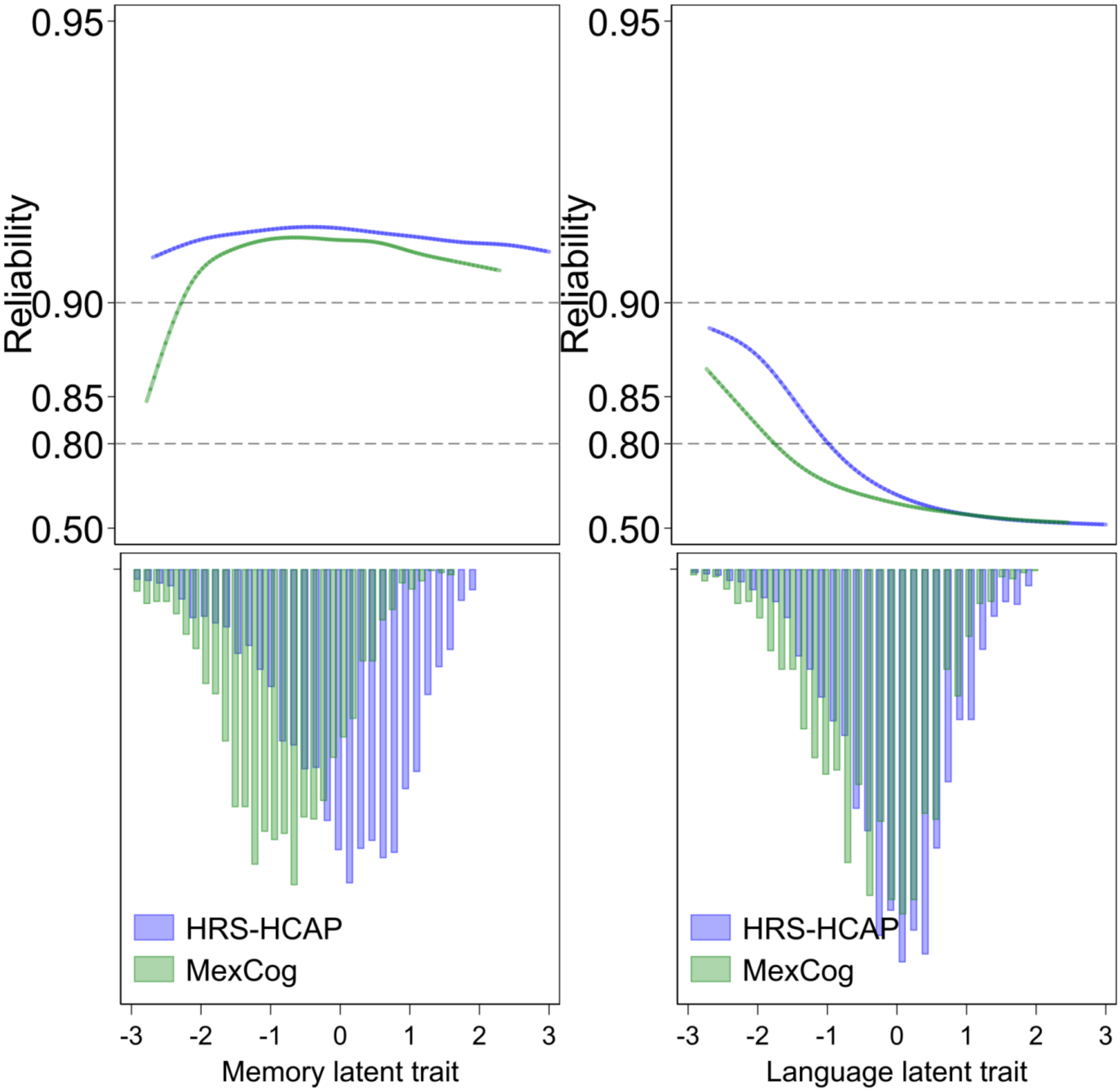
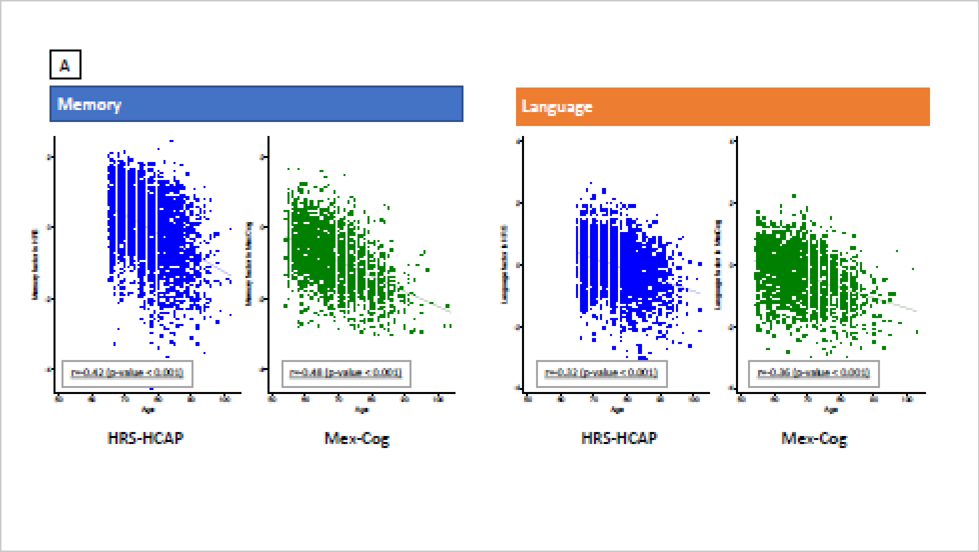


Figure 3. Panel A represents the associations between memory factor scores and age, and the associations between language factor scores and age by study. Panel B represents the associations between memory factor scores and years of education, and the associations between language factor scores and years of education by study.


A
Memory
Language
r=-0.42 (p-value < 0.001)
r=-0.48 (p-value < 0.001)
r=-0.32 (p-value < 0.001)
r=-0.36 (p-value < 0.001)
Age
Age
Age
Age
HRS-HCAP
Mex-Cog
HRS-HCAP
Mex-Cog