HHS Public Access

Author manuscript

Neuropsychology. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Neuropsychology. 2023 March ; 37(3): 247–257. doi:10.1037/neu0000816.

A cultural neuropsychological approach to harmonization of cognitive data across culturally and linguistically diverse older adult populations

Emily M. Briceño[1,*], **Miguel Arce Rentería**[2,*], **Alden L. Gross**[3], **Richard N. Jones**[4], **Christopher Gonzalez**[5], **Rebeca Wong**[6], **David R. Weir**[7], **Kenneth M. Langa**[8,9,10,11], **Jennifer J. Manly**[2]

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

[2]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

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

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

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

[6]Sealy Center on Aging, University of Texas Medical Branch at Galveston, Galveston, Texas, USA

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

[8]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI

[9]Institute for Social Research, University of Michigan, Ann Arbor, MI

[10]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor MI

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

Abstract

Objective: To describe a cultural neuropsychological approach to pre-statistical harmonization of cognitive data across the United States (US) and Mexico with the Harmonized Cognitive Assessment Protocol (HCAP).

Methods: We performed a comprehensive review of the administration, scoring, and coding procedures for each cognitive test item administered across the English and Spanish versions of the HCAP in the Health and Retirement Study (HRS) in the US and the Ancillary Study on

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.

*Co-first authors and co-corresponding authors

Cognitive Aging in Mexico (Mex-Cog). For items that were potentially equivalent across studies, we compared each cognitive test item for linguistic and cultural equivalence and classified items as confident or tentative linking items, based on the degree of confidence in their comparability across cohorts and language groups. We evaluated these classifications using differential item functioning techniques.

Results: We evaluated 132 test items among 21 cognitive instruments in the HCAP across the HRS and Mex-Cog. We identified 72 confident linking items, 46 tentative linking items, and 14 items that were not comparable across cohorts. Measurement invariance analysis revealed that 64% of the confident linking items and 83% of the tentative linking items showed statistical evidence of measurement differences across cohorts.

Conclusions: Pre-statistical harmonization of cognitive data, performed by a multidisciplinary and multilingual team including cultural neuropsychologists, can identify differences in cognitive construct measurement across languages and cultures that may not be identified by statistical procedures alone.

Keywords

harmonization; cognitive aging; methodology; cross-cultural; cultural neuropsychology

Introduction

There is increased interest in harmonization of cognitive data across culturally and linguistically diverse populations to better understand the biological, social, economic, and cultural factors contributing to cognitive aging and dementia. In order to adequately evaluate cognitive function and its sociocultural and health determinants across diverse populations, careful, precise, and culturally informed harmonization of the instruments is needed. However, optimal procedures for culturally informed harmonization of cognitive instruments have been understudied (Briceno et al., 2021).

There are many statistical procedures for harmonization of cognitive data, including Item Response Theory (IRT) based approaches that facilitate direct and quantitative comparisons across datasets collected in different contexts through co-calibration using both common and unique items across studies. While statistical approaches to harmonization of cognitive data have been described extensively (Griffith et al., 2013; Gross et al., 2014), less research is available on the steps required for pre-statistical harmonization of cognitive data (Briceno et al., 2021). The pre-statistical phase of cognitive data harmonization requires careful review of the neuropsychological instruments included in a given study by experienced neuropsychologists to determine whether the cognitive construct is measured consistently across studies. This work requires expertise in cognitive assessment, in addition to linguistic and cultural competence with the populations being harmonized. As such, cultural neuropsychologists are ideally suited to provide expertise in the pre-statistical and statistical harmonization of cognitive data across linguistically and culturally diverse populations.

The Harmonized Cognitive Assessment Protocol (HCAP) through the Health and Retirement Study (HRS) (Langa et al., 2020) is a flexible but comparable neuropsychological battery for measuring cognitive function among older adults around the world through the HRS' international partner studies, such as the Ancillary Study on Cognitive Aging in Mexico (Mex-Cog) (Mejia-Arango et al., 2020). Although the HRS-HCAP and Mex-Cog were designed to maximize comparability across cohorts, these studies have methodological, administration, and regional differences, which complicates direct comparison. Each study adapted the neuropsychological test battery to be culturally appropriate for their population, through adjustments such as modifying stimuli, translation of individual items/tests, as well as the administration and scoring procedures. No studies to date have offered procedural recommendations for content review of cross-cultural and cross-linguistic cognitive data to determine suitability for harmonization.

The present study sought to: 1) describe a cultural neuropsychological approach to pre-statistical harmonization of cognitive data across the US and Mexico; 2) describe procedures for evaluating measurement equivalence using item response theory methods; and 3) offer recommendations for future studies focused on cross-cultural cognitive data harmonization research in cognitive aging.

Methods

Cohorts:

HRS-HCAP cohort: The Harmonized Cognitive Assessment Protocol (HCAP) participants were recruited as a sub-study of the 2016 wave of the Health and Retirement Study (HRS) (Langa et al., 2020). The HRS is a nationally representative cohort study of contiguous US-dwelling adults aged 51 and older that has been ongoing biannually since 1992. HCAP ancillary study selection procedures are available elsewhere (Langa et al., 2020). Briefly, participants were selected from approximately half of the HRS sample who completed the 2016 core HRS interview and were aged 65 years or older. 3496 adults participated in the HCAP assessment, in English (95% of sample) or Spanish (5% of sample). The sample includes 383 Hispanic/Latinx participants, 551 non-Hispanic Black participants, 2483 non-Hispanic white participants, and 79 participants who identified as another race/ethnicity. All participants provided informed consent.

Mex-Cog: The Ancillary Study on Cognitive Aging in Mexico (Mex-Cog) began in 2016 as a sub-sample of the nationally representative Mexican Health and Aging Study (MHAS) (Mejia-Arango et al., 2020). MHAS is a nationally representative sample of Mexican adults age >=50, designed to prospectively evaluate the impact of disease on health, function, and mortality (Wong et al., 2015). Stratified sampling procedures were used to select a cohort of Mex-Cog participants that were 55 years and older, and from eight different Mexican states. The following criteria were used when selecting the eight states: socioeconomic factors (percent urban/rural, number of residents who returned after migrating to the United States) and health characteristics (percent with obesity, diabetes, mine industry, and pottery industry). A total of 2,265 participants were included in Mex-Cog, and 2,042 participants were administered the cognitive assessment. All participants provided informed consent.

Cognitive Assessment

The HCAP cognitive assessment battery was administered in the HRS-HCAP and Mex-Cog cohorts. Details regarding individual subtests comprising the HCAP are available elsewhere (Langa et al., 2020). As mentioned above, the HCAP battery was designed to be used across international population-based longitudinal studies of aging around the world. Generally, the HCAP battery, in both HRS-HCAP and Mex-Cog, includes measurement of various cognitive domains such as orientation, memory, language, visuospatial function, and executive functioning. Table 1 describes the neuropsychological subtests comprising the HCAP battery across the HRS-HCAP and Mex-Cog cohorts. Neuropsychological measures were grouped into cognitive domains as largely determined in prior published studies that have determined the factor structure of the HCAP battery (Arce Renteria et al., 2021; Gross et al., 2020; Jones et al., 2020).

Procedures

Review cohort characteristics: The University of Michigan Medical school IRB reviewed this study and determined it to be exempt. We first collected and compared characteristics of the cohorts that informed the degree to which the study populations could be considered comparable and to inform decision-making for determination of linking items across cohorts. Characteristics reviewed included sampling and recruitment procedures, geographic distribution of participants (e.g., rural/urban), educational characteristics, literacy, and languages used.

Review and documentation of cognitive instrument details: We obtained detailed information on all cognitive instruments administered across all studies, including both English and Spanish versions of the HRS-HCAP. While each study has published methodological documents (Mex-Cog, 2020), we requested access to the full batteries, codebooks, and any instruction and scoring manuals when available and permitted. We followed published pre-statistical harmonization procedures (Briceno et al., 2021). Briefly, we reviewed and documented information about each test, including test version, administration and scoring procedures, and study-specific adaptations (e.g., test stimulus adaptations). We carefully reviewed item translations for each test that was administered in English and Spanish. We documented possible (based on test structure) and observed score ranges for each test score.

Decision-making process for determining linking items: Determination of linking items required input from all members of the harmonization team, with representation from individuals with expertise in cultural neuropsychology, competence in the language and cultures represented in the original cohorts, the original data collection procedures, and in statistical harmonization procedures. After all previously described documentation was assembled, we carefully reviewed each item to determine its comparability across studies, or the degree to which the items may be interpreted equivalently and “linked” across studies. Comparability was assessed by the following factors: 1) review of all key procedural details to ensure equivalent administration and scoring, 2) review of any test adaptations to rule out study differences in test structure or implementation that could meaningfully impact test score interpretation, 3) evaluation/review of the translations of a given item to confirm

linguistic, cultural, and construct equivalence of the item from a theoretical perspective. We classified items as linking items if they were deemed equivalent across each of these comparability factors. Given the possibility of finding minor differences across these factors with an unclear impact on the equivalence of an item, we further classified linking items as *confident* linking items (i.e., no known factors violating item equivalence) and *tentative* linking items (i.e., possible factors violating item equivalence). Items that were determined to have factors that violated item equivalence were classified as unique items (i.e., not linking items).

Statistical procedures to detect differential item functioning in linking items: The validity of attempts to place estimates of cognitive functioning using data collected in different studies requires assumptions that some tasks or items are equivalent across study. The degree to which our expert panel has identified equivalent linking items can be evaluated with statistical tools for identifying items that behave differently across stable groups: differential item functioning (DIF) detection using item response theory methods (Camilli et al., 1994). We used an approach to DIF detection described previously (Jones, 2006) using the MIMIC model in Mplus (version 8.6, Muthén & Muthén, Los Angeles CA). Briefly, this method involves estimating categorical response variable factor analysis models with cognitive tasks or tests as the factor indicators, and a grouping variable (Mex-Cog vs HRS-HCAP) as a predictor of underlying (latent) cognition. Forwards stepwise model modifications using model misfit statistics (modification indices) are used to identify relationships between study membership and individual items (direct effects) that, if freely estimated, would significantly improve model fit. These direct effects are measures of DIF magnitude. We evaluate DIF impact by – after completing our DIF analysis – we estimate for each person their level on the latent trait ignoring DIF (i.e., assuming all items have the same measurement parameters across group) and again accounting for DIF (i.e., allowing the items identified with DIF to have different measurement model parameters across study). We then compare the difference in these two estimates at the individual participant level. This analysis step looks for evidence of *salient DIF* by computing the difference between non-DIF adjusted scores and DIF-adjusted scores. We calculate the proportion of participants with scores that differ by more than 0.3 SD units across the two sets of estimates (Goel & Gross, 2019).

We conducted DIF detection in two steps. First, we considered only those items identified as *confident linking items*. Any item identified with DIF that is greater than of negligible magnitude (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 (Zwick, 2012)) is removed from the confident item set. Following this analysis, we have a *DIF-free confident* linking item set and a *tentative* linking item set. The second analysis uses all items in the original *confident linking item* set and *tentative linking item* set, but the items in the *DIF-free confident* linking item set are treated as anchor items. Anchor items are items for which we assume there is no DIF. This assumption allows for model fitting to proceed with our assumptions about previously detected DIF, or the absence of DIF, to remain in place. See Jones et al (2019) for more discussion.

Transparency and openness: The present manuscript involves publicly available data and research materials through the HRS-HCAP and Mex-Cog websites, with the exception of copyrighted test material. We have described our sample size and all measures used in the study. Statistical analyses were conducted with MPlus. This study's design and its analysis were not pre-registered.

Results

Demographic characteristics of cohorts:

Table 2 displays demographic (age, sex/gender, and education) and cardiovascular (diabetes, stroke, hypertension) factors by cohort. The HRS/HCAP and MexCog samples did not differ with respect to sex or diabetes. But the HRS/HCAP was older (76.6 years vs 68.1 years, $p<0.001$), more highly educated ($p<0.001$), had more wealth ($p<0.001$), and had higher prevalence of stroke ($p<0.001$) and hypertension ($p<0.001$) compared to MexCog.

Summary of cognitive instrument items reviewed and linking items:

We reviewed 21 cognitive instruments across both the HCAP (English and Spanish) and Mex-Cog (Spanish) cognitive assessment batteries evaluating the domains of memory, language, visuospatial abilities, executive functioning, and orientation. Across these 21 tests, 132 items were available across the HRS-HCAP-English, HRS-HCAP-Spanish, and Mex-Cog. Each item was meticulously reviewed to determine comparability across cohorts. After review of test versions, cultural and linguistic considerations in translations, test administration, and scoring procedures, 72 items were classified as confident linking items across at least two cohorts, and an additional 46 items were classified as tentative linking items. The remaining 14 items were determined to be unique items within cohorts (i.e., not comparable) due to differences in the key comparability factors reviewed previously (Figure 1, Panel A). Furthermore, all individual linking items were grouped into pairs in order to compare them between HRS-HCAP-English and Mex-Cog for the DIF analyses (Figure 1, Panel B). Supplemental Table 1 describes all items, their linking status across cohorts, and rationale for the classification. We provide examples of these classifications and their rationale below.

Confident linking items.—Linking items were largely found among the cognitive domains of memory, language, visuospatial and orientation. Several items were clearly linking items due to nearly identical administration, scoring, and English-Spanish translations that did not appear to impact the meaning of the item (e.g., orientation to month, year, day of the week). However, certain items required recoding for classification as an item. For instance, all batteries included the MMSE item requiring participants to read and follow a command. However, there were administration and coding differences between HRS-HCAP and Mex-Cog, such that HRS-HCAP provided separate codes for reading and following the command correctly (coded as 1) and following the command correctly after being read the item (coded as 2). In Mex-Cog, the participants who could not read the item independently were not read the item, and instead received a missing data code. As such, we re-coded this item in the HCAP dataset (assigned those with codes of 2 to 'missing') to align with administration and coding procedures for Mex-Cog. To avoid the loss of this unique

information provided by the HRS-HCAP, we created a separate item for the HRS-HCAP (read and follow command, unable to read) in which codes of 2 (followed the command after being read the item) were retained, and other codes were coded as missing. Coding differences were also found for items with multiple steps. For example, the Community screening interview for dementia (CSI-D (Zhao et al., 2014)) item requiring following a 2-step command was administered in both HRS-HCAP and Mex-Cog. On this item, Mex-Cog coded two separate items for each correctly followed step, whereas HRS-HCAP provides one item, coding a point if both steps were followed correctly. As such, we re-coded the Mex-Cog data to align with that of HRS-HCAP. Lastly, some items were only classified as linking items between the English and Spanish version of the HRS-HCAP, as similar items were not included in Mex-Cog (e.g., orientation to address; Raven's progressive matrices).

Tentative linking items.—There were several items that we classified as tentative linking items based on minor concerns for non-equivalence based upon our review of the comparability factors. A key example of this was a scoring procedure difference for the WMS-IV Logical Memory test, included across all cohorts. In Mex-Cog, scoring of Logical Memory was altered to allow separate coding for both "exact" and "approximate" responses. The HRS-HCAP study followed standard WMS-IV scoring procedures according to the WMS-IV manual (Wechsler, 2009). Given that there were no additional details regarding this scoring procedure in the Mex-Cog methodological documents, we directly contacted key investigators for that study to obtain additional clarification. Through personal communication we were informed that the "exact" scoring in Mex-Cog aligns with the standard scoring procedures from the WMS-IV. Inspection of the Logical Memory test possible score range with the Mex-Cog "exact" scoring approach aligned with that of the HRS-HCAP possible score range (0–25). As such, we classified the Logical Memory scores using the "exact" scoring in Mex-Cog as a tentative linking item with HRS-HCAP.

Potential language differences related to item translations also presented as a source to confound whether an item could be considered a linking item across cohorts. For example, review of the individual items in the Recognition portion of WMS-IV Logical Memory revealed that two items in the Spanish HRS-HCAP appeared more difficult for the Spanish compared to the English version, whereas one item appeared more difficult in the English version compared to Spanish, all related to idiosyncrasies of the translation. For instance, in the Spanish version of WMS-IV Logical Memory a very common street name in Mexico is used in the story but a very uncommon street is provided as a foil in the recognition. Whereas, in the English version of the test, both the target and foil street names are similarly familiar. As such, this test was classified as a tentative linking item between the English and Spanish versions of the HRS-HCAP. The recognition portion of WMS-IV Logical Memory is not administered in Mex-Cog.

Lastly, beyond issues related to language and translation, cultural differences in approach and performance on cognitive tests were considered. For example, the Trail Making Test Part B was classified as a tentative linking item across the English and Spanish HRS-HCAPs due to concerns for cultural differences impacting test interpretation. Studies suggest that individuals from Spanish-speaking backgrounds perform systematically slower on Trails B compared to English-speakers (Acevedo et al., 2007; Kisser et al., 2012). Differences in

performance among English and Spanish-speakers has been associated with sociocultural factors such as degree of acculturation which impact familiarity with the test items and differences in cultural values regarding speed and accuracy (Acevedo et al., 2007; Rosselli & Ardila, 2003). Given these issues, we classified Trails B as a tentative linking item across English and Spanish HRS-HCAP.

Non-linking items.—Some items were easy to determine as non-linking (i.e., unique) items due to simply not having a similar test among the three cohorts (i.e., Go/No-go in Mex-Cog). Some potentially similar tests were determined to be non-linking items, such as the Symbols and Digits test, due to significant differences in test stimuli and administration that impacted score interpretation. In Mex-Cog, the test is the Symbols and Digits test, in which participants must fill in the correct symbol for a given number based on a key provided. This test includes 56 items and has a time limit of 90 seconds. However, the HRS-HCAP version of the test is the Symbol Digit Modalities Test (SDMT (Smith, 1982)), which requires the participant to enter a number for a given symbol based on a key; it also has a time limit of 90 seconds but includes 110 items. Given these differences in stimuli and administration this was considered as non-comparable with Mex-Cog. In addition, given the potential cultural differences associated with the construct validity of the SDMT across English- and Spanish-speaking individuals (Arango-Lasprilla et al., 2015; O'Bryant et al., 2007), we classified it as a tentative linking item between the HRS-HCAP English and Spanish batteries.

Another test that may appear comparable if not carefully scrutinized was the CERAD Word List Memory Test (Morris et al., 1989). While all studies included the CERAD, the administration differed notably between HRS-HCAP and Mex-Cog. The HRS-HCAP study presented the test stimuli (list of words) visually and asked the participant to read the words aloud, whereas in Mex-Cog the words are presented verbally by the examiner to the participant, with no visual stimuli. In addition, the order of the list of words is randomized in each trial in HCAP but the lists were presented in the same order across all three trials in Mex-Cog. Given the literature that supports differences in performance between verbal and visual presentation of memory tests (Constantinidou & Baker, 2002; Fougnie & Marois, 2011) and the impact of variable ordering of items on test performance (Gross & Rebok, 2011), we determined that these administration differences affected score interpretation for all items (i.e., total list recall, delayed recall, recognition) between HRS-HCAP and Mex-Cog. However, given that these differences did not vary between the English and Spanish versions of the HRS-HCAP, we did consider the items as tentative linking items.

DIF statistical harmonization results.—For the following analyses, we evaluated DIF between HRS-HCAP English with Mex-Cog, excluding the Spanish version of the HRS-HCAP. Given that the Spanish HRS-HCAP sample size is not large enough for reliable DIF analyses between two cohorts (N = 178; 5% of HRS-HCAP sample), for a cleaner DIF analysis comparing the HRS-HCAP to Mex-Cog, the remainder of the DIF analyses compared English-speaking participants in HRS-HCAP with participants in Mex-Cog (i.e., excluding Spanish speaking participants in HRS-HCAP).

Confident linking items—First, we evaluated for differential item functioning among the 14 confident linking item pairs between HRS-HCAP and Mex-Cog. DIF analyses revealed that of these 14 linking item pairs, 5 item pairs (36%) were found to be measuring the underlying construct in a similar fashion in both HRS-HCAP and Mex-Cog, whereas 9 item pairs (64%) showed evidence of DIF (Table 3). Three of these 9 items showed DIF of greater than negligible magnitude, all items from the MMSE: orientation to year, orientation to state, and reading and following a command. We next evaluated whether these items with DIF had a meaningful impact on the factor scores by examining whether the difference between DIF-adjusted scores and non-DIF adjusted scores exceeded a threshold of 0.3 SD units (Goel & Gross, 2019). We found no evidence of meaningful (salient) DIF on the factor score generated from this confident linking item set (Figure 2, Panel A).

Tentative items allowed to demonstrate DIF—Second, we evaluated for DIF in our tentative linking item set. In this analysis, we included the confident linking items that showed either no DIF or negligible DIF in the first step of our analysis as anchor items for this analysis (3 confident linking items removed, leaving 11 of our 14 confident linking items as anchor items for DIF analysis). This analysis thus tested for DIF among the 12 tentative linking item pairs and 3 confident linking items pairs that were identified as having non-negligible DIF in step 1. This analysis revealed that of the 12 tentative linking item pairs, 2 item pairs (17%) were found to be measuring the underlying construct in a similar fashion in both HRS-HCAP and Mex-Cog, whereas 10 item pairs (83%) showed evidence of DIF (Table 3). Among these item pairs, only 4 were indicative of negligible DIF (MMSE 3-word delay, Logical Memory Delay, Brave Man immediate recall, and the MMSE write a sentence), all other items showed DIF of greater than negligible magnitude. Considering salient DIF, only n = 1 (<1%) of HRS-HCAP participants' non-DIF-adjusted scores were = > 0.3 SD units different than their DIF-adjusted scores, whereas a greater proportion (N=803, 39%) of Mex-Cog participants' non-DIF-adjusted scores were = > 0.3 SD units different than their DIF-adjusted scores. Said another way, not accounting for DIF would lead to a considerable depression in scores among Mex-Cog participants.

Discussion

The present manuscript describes a cultural neuropsychological approach to harmonization of cognitive data across linguistically and culturally diverse older adults. To harmonize cognitive data from the Harmonized Cognitive Assessment Protocol (HCAP) across the Health and Retirement Study (HRS) and Mexico (Mex-Cog), we performed a comprehensive review of each cognitive test item administered in each cohort. We evaluated each potentially comparable item through a cultural neuropsychological lens to consider whether each item exhibited conceptual equivalence across cohorts. We classified items as tentative or confident linking items, based upon this comprehensive review. Although we found statistical evidence for measurement non-equivalence among several (9 of 14) of our confident linking items across cohorts, the majority of these items (6 of 9) were of negligible magnitude and did not substantially impact the measurement of cognition across these groups. This multidisciplinary approach to harmonizing cognitive data across languages

and cultures is necessary for appropriate inferences about cognitive health in culturally and linguistically diverse populations.

We found small cross-cohort measurement differences across several items that we had initially assumed to be equivalent (i.e., confident linking items) across cohorts, based on our careful assessment of construct equivalence from a theoretical perspective. These items included several items from the MMSE and two items from the 10/66 dementia assessment. One possible explanation for some of these measurement differences may be undocumented differences in scoring procedures across cohorts. Certain details regarding scoring procedures were not available in documentation; as such, it is possible that there may have been study-specific nuances in degree of leniency in determination of acceptable responses. Although there were several items that we had labeled as confident yet showed to measure the cognitive construct differently across cohorts, the impact of DIF on the latent factor scores was small. In contrast, we found meaningful measurement differences (of at least 0.3 standard deviations) among our tentative linking items due to theoretical concerns of possible non-equivalence, which resulted in a meaningful underestimation of harmonized factor scores in 39% of the Mex-Cog sample.

The assignment of tentative linking items, with documentation of rationale for possible measurement differences across cohorts, facilitates interpretation of statistical measurement invariance analyses. This a priori review and documentation can facilitate decision-making regarding integration of statistical evidence of measurement difference with pre-statistical linking item decisions. For example, we classified the MMSE phrase repetition as a tentative linking item due to concerns that it may be a more common phrase in Spanish in Mexico, making the item possibly easier. We also noted that the scoring of the English phrase was more stringent, also possibly leading to greater error rates on this item in English. Our DIF findings aligned with this concern identified during pre-statistical harmonization, which strengthens the scientific rationale for assigning it as a unique item rather than a linking item for subsequent generation of harmonized factor scores.

Statistical harmonization offers promise in directly comparing cognition across studies, including cross-national and cross-linguistic comparisons. The IRT approach to statistical harmonization allows for co-calibration of cognitive measures across groups, even in the context of differences in language of administration and item content, given that the assumption of conceptual equivalence is met and at least some cognitive tests overlap across studies (Chan et al., 2015; Eremenco et al., 2005). Pre-statistical harmonization is a critical step of this process, to ensure sufficient conceptual equivalence and sufficient equivalence in overlapping cognitive tests across studies. However, few studies have focused on pre-statistical procedures for harmonization of cognitive data (Briceno et al., 2021) and prior work has shown that details on harmonization procedures are rarely described (Griffith et al., 2013). Although general best practices in cognitive data harmonization apply across all circumstances, cross-cultural and cross-national cognitive data harmonization endeavors offer unique challenges and considerations. Determining equivalence in cognitive test information across languages and cultures is an understudied and complex endeavor in neuropsychology (Fernández & Abe, 2018; Pedraza & Mungas, 2008). If these nuances and complexities are not carefully considered, there is risk for misinterpretation of cross-cultural

differences in cognitive test scores that may serve as barriers to accurate attributions of the determinants and impact of cognitive health. Cultural neuropsychology expertise is critical to the harmonization process, particularly in guiding the pre-statistical linking item decisions. When done correctly, the derived harmonized factor scores from each study can then be meaningfully compared to examine the socio-cultural mechanisms underlying cognitive aging inequalities across a range of diverse populations.

Our findings underscore the importance of transparency in methodology and decision-making about pre-statistical harmonization decisions and their scientific rationale. For items that show some evidence for non-equivalence across cohorts, maintaining them as linking items may benefit the stability of the harmonized factor scores, given that more linking items are available (Gross et al., 2014). However, the presence of DIF in linking items may bias the estimation of factor scores and result in disproportionate measurement error across groups. As such, it is important to integrate available information, including pre-statistical and statistical evidence of measurement non-equivalence in order to assign linking items for a given study. These decisions should also be made transparently to allow for scientific discourse to advance the science of this methodology.

In order to aid future cross-cultural cognitive harmonization research, we provide the following recommendations regarding cognitive data harmonization teams. The key areas of expertise needed for cross-cultural and cross-linguistic cognitive data harmonization procedures included a clinical neuropsychologist with expertise in cultural neuropsychology, a statistician with expertise in statistical harmonization methods such as item response theory, study team members from the individual cohort studies to advise on any procedures that were not documented in study procedures, and individuals with cultural and linguistic competence in the languages and cultures represented in the cohort studies. The documentation required for the pre-statistical phase of harmonization can be onerous and the most time-intensive task of the pre-statistical harmonization work. In our experience, 8–10 hours of documentation is needed per 1 hour of neuropsychological test battery. The neuropsychologist's additional work including the review, integration, and determination of linking items requires approximately 4 hours per 1 hour of neuropsychological test battery. As such, to aid with the documentation procedures described in this manuscript, a research assistant with linguistic competence in the languages used for assessment in addition to experience with cognitive test administration may be included in the harmonization team. Linguistic competence and cognitive test administration experience is necessary for the efficient extraction of the relevant details regarding test administration and scoring. We recruited a bilingual (English/Spanish) graduate student in clinical neuropsychology for this role; a psychometrist or a research assistant with prior neuropsychological testing experience may also have the appropriate skill set for this role. When deciding the pre-statistical harmonization linking item designations, it was beneficial to have a consensus approach in determining linking item designation versus relying on one single expert. Our team benefited from the inclusion of two neuropsychologists with expertise cultural neuropsychology and harmonization research (MAR, EB) that took a consensus approach when evaluating all comparability factors to inform linking designations. Future studies are recommended to adopt a similar approach to pre-statistical harmonization.

Our study has implications for future research. Although cognitive test harmonization is rapidly growing as interest grows in leveraging large population-based datasets for investigation of biomedical and sociocultural contributors to cognitive health across the lifecourse, future research is needed to better determine best practices in cross-cultural cognitive data harmonization, such as transparency in the procedures used and decisions made throughout the harmonization process, and integration of the pre-statistical and statistical phases of harmonization to determine appropriate linking items. Although researchers are often faced with determination of linking items and interpreting DIF findings themselves, community members represented in the cohort studies may offer invaluable observations regarding the cultural and linguistic relevance of test items, which may be useful to inform decision making about cross-cultural equivalence in the approach to answering cognitive test items. Future work may utilize procedures such as cognitive interviewing as a methodology to capture this unique information. In addition, it will be important to examine the extent to which harmonized cognitive test scores relate to indicators of brain health and to daily functioning, which may also vary across cultures. Finally, we reported cohort differences in some demographic and cardiovascular health characteristics, which was an expected finding given the known characteristics of these populations. Future research is needed to further evaluate the various potential factors contributing to the observed measurement differences across cohorts.

There are a few limitations to the current study that warrant mention. First, this work was completed as a secondary data analysis on previously collected data. Given that we were not directly involved with either study from the outset, we had to rely on available documentation and information provided through personal communication with study staff. As such, although both studies provided detailed and careful documentation, we cannot rule out the possibility of errors due to unknown, undocumented variability in data collection procedures. While harmonization of cognitive data from cohort studies may be facilitated by inclusion of neuropsychologists at the study design phase, the approach highlighted in our manuscript is likely applicable when harmonizing independent cohort studies with cognitive outcomes. Second, our approach may not be generalizable to all cohorts of differing language and cultural backgrounds. Our approach carefully considered issues related to differences between English and Spanish and potential cultural differences associated with Latin American cultures, specifically Mexican cultural background. There may be unique considerations for harmonization for other languages and cultures that were not represented in our work. For instance, an HRS international partner study that administers a version of the HCAP cognitive battery is the Longitudinal Aging Study in India- Diagnostic Assessment of Dementia (LASI-DAD)(Lee et al., 2020). The LASI-DAD currently is administered in 13 languages across India, and these linguistic factors require their own unique pre-statistical harmonization considerations. Similarly, if working with studies such as the China Health and Retirement Longitudinal Study (CHARLS) (Zhao et al., 2014), in which there are structural differences in reading and writing between English and Mandarin, new and different harmonization considerations may arise. However, we believe that our approach and recommendations, particularly the multidisciplinary, multicultural, and multilingual characteristics of the harmonization team, can serve as a foundation for

investigators working with culturally and linguistically diverse cohorts to include in their harmonization efforts.

Statistical harmonization of cognitive data allows direct comparison of cognitive functioning between different cohorts to enhance our understanding of factors of risk and resilience on cognition. Cognitive data harmonization across culturally and linguistically diverse populations requires careful procedures and multidisciplinary expertise to optimize construct equivalence across groups. We advocate for a cultural neuropsychological approach to harmonization that integrates theoretical and statistical evidence for measurement equivalence across diverse populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

This research was completed with funding from the NIH/HHS (5-R24-AG-065182-02, EB, MAR, KL, DW), and NIA (1K99AG066932-01, MAR). 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

References

Acevedo A, Loewenstein DA, Agrón J, & Duara R (2007). Influence of sociodemographic variables on neuropsychological test performance in Spanish-speaking older adults. Journal of Clinical and Experimental Neuropsychology, 29(5), 530–544. [PubMed: 17564918]

Arango-Lasprilla J, Rivera D, Rodríguez G, Garza M, Galarza-Del-Angel J, Rodriguez W, Velazquez-Cardoso J, Aguayo A, Schebela S, & Weil C (2015). Symbol digit modalities test: normative data for the Latin American Spanish speaking adult population. NeuroRehabilitation, 37(4), 625–638. [PubMed: 26639927]

Arce Renteria M, Manly JJ, Vonk JMJ, Mejia Arango S, Michaels Obregon A, Samper-Ternent R, Wong R, Barral S, & Tosto G (2021, Aug 11). Midlife Vascular Factors and Prevalence of Mild Cognitive Impairment in Late-Life in Mexico. J Int Neuropsychol Soc, 1–11. 10.1017/S1355617721000539

Briceno EM, Gross AL, Giordani BJ, Manly JJ, Gottesman RF, Elkind MSV, Sidney S, Hingtgen S, Sacco RL, Wright CB, Fitzpatrick A, Fohner AE, Mosley TH, Yaffe K, & Levine DA (2021). Pre-Statistical Considerations for Harmonization of Cognitive Instruments: Harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS. J Alzheimers Dis, 83(4), 1803–1813. 10.3233/JAD-210459 [PubMed: 34459397]

Camilli G, Shepard LA, & Shepard L (1994). Methods for identifying biased test items (Vol. 4). Sage.

Chan KS, Gross AL, Pezzin LE, Brandt J, & Kasper JD (2015, Dec). Harmonizing Measures of Cognitive Performance Across International Surveys of Aging Using Item Response Theory. J Aging Health, 27(8), 1392–1414. 10.1177/0898264315583054 [PubMed: 26526748]

Constantinidou F, & Baker S (2002). Stimulus modality and verbal learning performance in normal aging. Brain and language, 82(3), 296–311. [PubMed: 12160526]

Eremenco SL, Cella D, & Arnold BJ (2005, Jun). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof, 28(2), 212–232. 10.1177/0163278705275342 [PubMed: 15851774]

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

Fougnie D, & Marois R (2011). What limits working memory capacity? Evidence for modality-specific sources to the simultaneous storage of visual and auditory arrays. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(6), 1329. [PubMed: 21859231]

Goel A, & Gross A (2019). Differential item functioning in the cognitive screener used in the Longitudinal Aging Study in India. International psychogeriatrics, 31(9), 1331–1341. [PubMed: 30782222]

Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, & Belleville S (2013). In Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. https://www.ncbi.nlm.nih.gov/pubmed/23617017

Gross AL, Jones RN, Fong TG, Tommet D, & Inouye SK (2014). Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology, 42(3), 144–153. 10.1159/000357647 [PubMed: 24481241]

Gross AL, Khobragade PY, Meijer E, & Saxton JA (2020, Aug). Measurement and Structure of Cognition in the Longitudinal Aging Study in India-Diagnostic Assessment of Dementia. J Am Geriatr Soc, 68 Suppl 3, S11–S19. 10.1111/jgs.16738 [PubMed: 32815599]

Gross AL, & Rebok GW (2011). Memory training and strategy use in older adults: results from the ACTIVE study. Psychology and aging, 26(3), 503. [PubMed: 21443356]

Jones JNMJJ, Langa KM, Ryan LH, Levine DA, McCammon R, & Weir D (2020). Factor structure of the Harmonized Cognitive Assessment Protocol neuropsychological battery in the Health and Retirement Study. PsyArXiv 10.31234/osf.io/rvmhj

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, S124–S133. [PubMed: 17060819]

Jones RN (2019). Differential item functioning and its relevance to epidemiology. Current epidemiology reports, 6(2), 174–183. [PubMed: 31840016]

Kisser JE, Wendell CR, Spencer RJ, & Waldstein SR (2012). Neuropsychological performance of native versus non-native English speakers. Archives of Clinical Neuropsychology, 27(7), 749–755. [PubMed: 22985952]

Langa KM, Ryan LH, McCammon RJ, Jones RN, Manly JJ, Levine DA, Sonnega A, Farron M, & Weir DR (2020). The Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: Study Design and Methods. Neuroepidemiology, 54(1), 64–74. 10.1159/000503004 [PubMed: 31563909]

Lee J, Khobragade PY, Banerjee J, Chien S, Angrisani M, Perianayagam A, Bloom DE, & Dey AB (2020). Design and Methodology of the Longitudinal Aging Study in India-Diagnostic Assessment of Dementia (LASI-DAD). Journal of the American Geriatrics Society, 68, S5–S10. [PubMed: 32815602]

Mejia-Arango S, Nevarez R, Michaels-Obregon A, Trejo-Valdivia B, Mendoza-Alvarado LR, Sosa-Ortiz AL, Martinez-Ruiz A, & Wong R (2020, Jul 27). The Mexican Cognitive Aging Ancillary Study (Mex-Cog): Study Design and Methods. Arch Gerontol Geriatr, 91, 104210. 10.1016/j.archger.2020.104210 [PubMed: 32781379]

Mex-Cog. (2020). Study on Cognitive Aging Linked to MHAS: Methodological Document, Version 1, November 2018

Morris JC, Heyman A, Mohs RC, Hughes J, van Belle G, Fillenbaum G, Mellits E, & Clark C (1989). The consortium to establish a registry for Alzheimer's disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology.

O'Bryant SE, Humphreys JD, Bauer L, McCaffrey RJ, & Hilsabeck RC (2007). The influence of ethnicity on Symbol Digit Modalities Test performance: An analysis of a multi-ethnic college and hepatitis C patient sample. Applied Neuropsychology, 14(3), 183–188. [PubMed: 17848129]

Pedraza O, & Mungas D (2008). Measurement in cross-cultural neuropsychology. Neuropsychology review, 18(3), 184–193. [PubMed: 18814034]

Rosselli M, & Ardila A (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. Brain and cognition, 52(3), 326–333. [PubMed: 12907177]

Smith A (1982). Symbol digit modalities test. Western Psychological Services Los Angeles.

Wechsler D (2009). WMS-IV: Wechsler memory scale. Pearson.

Wong R, Michaels-Obregon A, Palloni A, Gutierrez-Robledo LM, Gonzalez-Gonzalez C, Lopez-Ortega M, Tellez-Rojo MM, & Mendoza-Alvarado LR (2015). Progression of aging in Mexico: the Mexican Health and Aging Study (MHAS) 2012. Salud Publica Mex, 57 Suppl 1, S79–89. 10.21149/spm.v57s1.7593 [PubMed: 26172238]

Zhao Y, Hu Y, Smith JP, Strauss J, & Yang G (2014). Cohort profile: the China health and retirement longitudinal study (CHARLS). International journal of epidemiology, 43(1), 61–68. [PubMed: 23243115]

Zwick R (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, 2012(1), i–30.

Key points:

- Question: To combine information collected about cognitive health of older adults across different countries, it is important to consider whether cognition is measured equivalently across culturally and linguistically diverse cohorts. We described a cultural neuropsychological approach to harmonization of cognitive data across culturally and linguistically diverse cohorts in the United States and Mexico.
- Findings: We identified a set of cognitive test items that showed equivalent measurement across studies, and several cognitive test items that showed measurement differences.
- Importance: Comprehensive and careful comparison of cognitive data through the lens of a cultural neuropsychologist is needed for appropriate combination of cognitive data across countries.
- Next steps: Future studies may adapt and expand this approach given specific cultural and linguistic factors associated with cohorts outside of the United States and Mexico.

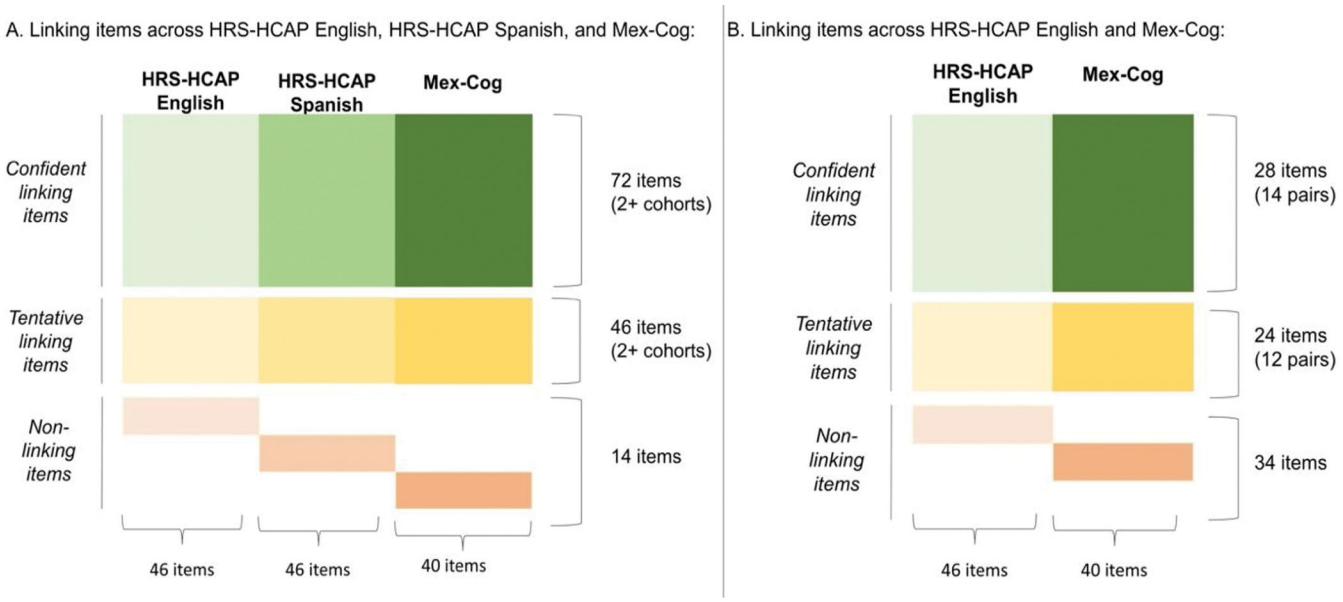

Figure 1.
Linking items and confidence ratings across cohorts.

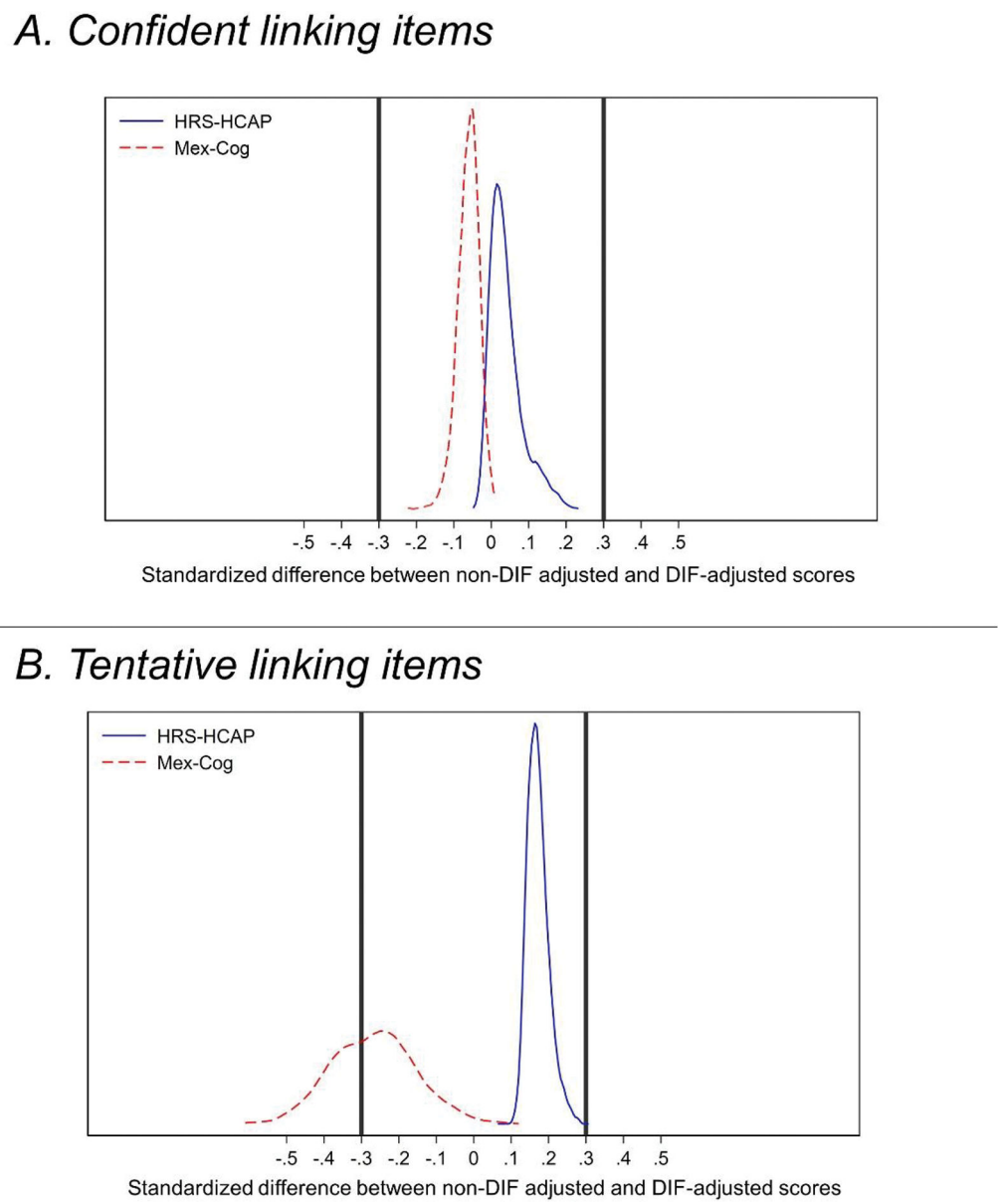

Figure 2.

Comparison of DIF-adjusted and non-DIF-adjusted General Cognitive Performance factor scores in HRS-HCAP and Mex-Cog among the confident and tentative linking items.

Figure Caption: Kernel density plots of standardized differences between non-DIF-adjusted factor scores and factor scores incorporating DIF. Y axis displays the probability density function describing the relative proportion of observations, stratified by cohort, along the x-axis. Panel A displays factor scores generated from confident linking items only. Panel B displays factor scores generated from both confident and tentative linking items, with confident linking items that showed negligible or no significant DIF treated as DIF-free anchor items.

Table 1.

Neuropsychological measures across HRS-HCAP and Mex-Cog

Cognitive Domain	Test	HRS-HCAP (English and Spanish)	Mex-Cog (Spanish)
Various	Community Screening Instrument for Dementia (CSI-D)	X	X
	MMSE	X	X
Language/Fluency	HRS-TICS	X	
	Semantic Fluency (Animal Naming)	X	X
Memory	CERAD Word List Learning and Recall-Immediate	X	X
	CERAD Word List Learning and Recall- Delayed	X	X
	CERAD Word List- Recognition	X	X
	WMS-IV Logical Memory-Immediate (Story B)	X	X
	WMS-IV Logical Memory (Story B)-Delayed	X	X
	WMS-IV Logical Memory (Story B)-Recognition	X	
	East Boston Memory Test-Immediate	X	X
	East Boston Memory Test- Delayed	X	X
	CERAD Constructional Praxis-Delayed Recall	X	X
Visuospatial	CERAD Constructional Praxis-Copy	X	X
Attention/Executive functioning	HRS Number Series	X	
	Letter Cancellation Test	X	
	Raven's Standard Progressive Matrices	X	
	Symbols and Digits[1]	X	X
	Trail Making Test (Parts A and B)	X	
	Go no go		X
	Serial 3 subtractions		X
	Serial 7 subtractions		X
	Similarities		X
	Visual Scan		X
	Timed Backward Counting Test	X	X

Note: CERAD is Consortium to Establish a Registry for Alzheimer's Disease. HRS is Health and Retirement Study. MMSE is Mini Mental State Examination. TICS is Telephone Interview for Cognitive Status. WMS-IV is Wechsler Memory Scale-Fourth Edition.

[1]Symbols and Digits was the Symbol Digit Modalities Test (SDMT) for HRS.

Table 2.

Sociodemographic and health characteristics of participants

Sample Characteristic	HRS-HCAP (n = 3496)	MexCog (n = 2042)	*p*
Age (*M, SD*)	76.6 (7.5)	68.1 (9.0)	< 0.001
Sex/Gender (*n*, column % female)	2095 (59.9)	1203 (58.9)	0.459
Education (*n*, column %)			
No Formal Education	22 (0.6)	350 (17.3)	< 0.001
Early Childhood Education	0 (0.0)	673 (33.3)	
Primary School (Grade 1–5)	158 (4.5)	452 (22.3)	
Lower Secondary (Grades 6–8)	251 (7.2)	317 (15.7)	
Upper Secondary (Grades 9–12)	1445 (41.4)	60 (3.0)	
Some College	765 (21.9)	156 (7.7)	
College or more	850 (24.3)	16 (0.8)	
History of Stroke (*n*, column %)	342 (9.9)	82 (4.0)	< 0.001
Type 2 Diabetes (*n*, column %)	1029 (29.7)	575 (28.2)	0.219
Hypertension (*n*, column %)	973 (28.4)	894 (43.8)	< 0.001

Table 3.

DIF results among confident and tentative linking items.

Step	Cognitive test item	Association with cohort[1]		
DIF among confident items		**Odds Ratio**	**95% CI**	**Interpretation[2]**
	Orientation-year	1.88	1.70, 2.07	DIF
	Orientation-state	2.86	2.44, 3.34	DIF
	Read and follow command (MMSE)	1.52	1.35, 1.71	DIF
	Orientation-day of month	1.22	1.13, 1.32	Negligible
	3-step command (MMSE)	1.14	1.05, 1.23	Negligible
	Orientation-month	1.08	0.96, 1.22	Negligible
	Follow 2-step command (10/66)	1.27	1.09, 1.49	Negligible
	Name (writing utensil; MMSE)	1.26	1.03,1.53	Negligible
	Name elbow (10/66)	0.95	0.80, 1.13	Negligible
DIF among tentative items, treating DIF-free confident items as anchors				
	Repetition of phrase (standard MMSE phrase)	0.32	0.29, 0.34	DIF
	Where is the local market	1.75	1.62, 1.89	DIF
	Logical Memory immediate	1.63	1.54, 1.73	DIF
	What do you do with a hammer	0.50	0.43, 0.57	DIF
	3-word immediate	0.61	0.56, 0.68	DIF
	3-word delay	1.31	1.24, 1.39	Negligible
	Logical Memory delay	1.31	1.24,1.39	Negligible
	Brave man immediate	1.20	1.13, 1.27	Negligible
	TICS Name scissors	0.57	0.46, 0.69	DIF
	MMSE write a sentence	1.06	0.96, 1.17	Negligible
Previously confident items that show DIF alongside tentative items				
	Orientation-What state are we in	2.86	2.44, 3.34	DIF
	Orientation-year	1.88	1.70, 2.07	DIF
	MMSE read and follow command	1.52	1.35, 1.71	DIF

Note:

[1]Reference group is Mex-Cog.

[2]Interpretation of the magnitude of DIF as negligible (between 0.66–1.5) or non-negligible (DIF). The beta coefficient is the difference (on a log odds scale) in outcome between HRS-HCAP and MexCog, adjusting for the latent ability. A positive coefficient implies better performance than expected on the item in HRS-HCAP, compared to MexCog, while a negative coefficient indicates better performance on the item than expected in MexCog, compared to HRS-HCAP.