## Title

Pre-Statistical Considerations for Harmonization of Cognitive Instruments: Harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS.



## Journal

Journal of Alzheimer's Disease, 83(4)

## ISSN

1387-2877

## Authors

Briceño, Emily M
Gross, Alden L
Giordani, Bruno J
et al.

## Publication Date

2021

## DOI

10.3233/jad-210459

Peer reviewed

HHS Public Access

Author manuscript

*J Alzheimers Dis*. Author manuscript; available in PMC 2022 January 06.

Published in final edited form as:

*J Alzheimers Dis*. 2021 ; 83(4): 1803–1813. doi:10.3233/JAD-210459.

# Pre-statistical Considerations for Harmonization of Cognitive Instruments: Harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS

**Emily M. Briceño, PhD**[a,b,c], **Alden L. Gross, PhD MHS**[d], **Bruno J. Giordani, PhD**[e,f], **Jennifer J. Manly, PhD**[g,h], **Rebecca F. Gottesman, MD, PhD**[i], **Mitchell S.V. Elkind, MD, MS**[g,j], **Stephen Sidney, MD, MPH**[k], **Stephanie Hingtgen, MPP**[l], **Ralph L. Sacco, MD, MS**[m], **Clinton B. Wright, MD, MS**[n], **Annette Fitzpatrick, PhD, MA**[o], **Alison E. Fohner, PhD, MS**[p], **Thomas H. Mosley, PhD**[q], **Kristine Yaffe, MD**[r], **Deborah A. Levine, MD, MPH**[b,c,l,s]

[a]Department of Physical Medicine and Rehabilitation, University of Michigan Medical School, 325 E. Eisenhower Blvd, Ann Arbor, MI, 48108

[b]Cognitive Health Services Research Program, University of Michigan Medical School, 2800 Plymouth Road, Bldg. 16, Ann Arbor, MI 48109

[c]Institute for Healthcare Policy and Innovation, University of Michigan, 2800 Plymouth Road, Bldg. 16, Ann Arbor, MI 48109

[d]Department of Epidemiology, Johns Hopkins Bloomberg School Public Health, 2024 E. Monument Street, Baltimore, MD 21205

[e]Departments of Psychiatry, Neurology, Psychology, and School of Nursing; University of Michigan, 2101 Commonwealth Blvd, Suite C, Ann Arbor, MI 48105

[f]Mary A. Rackham Institute, University of Michigan, 500 E Washington St #100, Ann Arbor, MI 48104

[g]Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University, 630 West 168th Street, P&S Box 16, New York, NY 10032

[h]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, 630 West 168th Street, P&S Box 16, New York, NY 10032

[i]National Institute of Neurological Disorders and Stroke Intramural Research Program, Bethesda, MD. Disclaimer: This article was prepared while Dr. Rebecca Gottesman was employed at the Johns Hopkins University School of Medicine. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States Government.

[j]Department of Epidemiology, Mailman School of Public Health, Columbia University, 710 West 168th Street, New York, NY 10032

Corresponding author: Emily M. Briceño, Ph.D., University of Michigan Medical School, Department of Physical Medicine and Rehabilitation, Rehabilitation Psychology and Neuropsychology, 325 East Eisenhower Parkway, Ann Arbor, MI 48108, Phone: 734-763-6496, emilande@med.umich.edu.

[k]Kaiser Permanente Northern California Division of Research, 2000 Broadway, Oakland, CA 94612

[l]Department of Internal Medicine, University of Michigan (U-M), 2800 Plymouth Road, Bldg. 16

[m]Department of Neurology, University of Miami, 1120 NW 14th Street, Suite 1352 Miami, FL 33136

[n]Division of Clinical Research, National Institute of Neurological Disorders and Stroke (NINDS), Bethesda, MD

[o]Department of Epidemiology, University of Washington, Office F-358A, Health Sciences Building, Box: 354696, Seattle, WA 98195

[p]Department of Epidemiology, University of Washington, Health Sciences Building, F-247A, Box 357236, Seattle, WA 98195

[q]Department of Medicine-Geriatrics, University of Mississippi Medical Center, 2500 N. State St., Jackson, Mississippi 39216

[r]Departments of Psychiatry, Neurology and Epidemiology, University of California, San Francisco, 4150 Clement St, San Francisco, CA 94121

[s]Department of Neurology and Stroke Program, University of Michigan, 2800 Plymouth Road, Bldg. 16, Room 430W, Ann Arbor, MI 48109-2800

## Abstract

**BACKGROUND:** Meta-analyses of individuals' cognitive data are increasing to investigate the biomedical, lifestyle, and sociocultural factors that influence cognitive decline and dementia risk. Pre-statistical harmonization of cognitive instruments is a critical methodological step for accurate cognitive data harmonization, yet specific approaches for this process are unclear.

**OBJECTIVE:** To describe pre-statistical harmonization of cognitive instruments for an individual-level meta-analysis in the blood pressure and cognition (BP COG) study.

**METHODS:** We identified cognitive instruments from six cohorts (the Atherosclerosis Risk in Communities Study, Cardiovascular Health Study, Coronary Artery Risk Development in Young Adults study, Framingham Offspring Study, Multi-Ethnic Study of Atherosclerosis, and Northern Manhattan Study) and conducted an extensive review of each item's administration and scoring procedures, and score distributions.

**RESULTS:** We included 153 cognitive instrument items from 34 instruments across the six cohorts. Of these items, 42% were common across ≥2 cohorts. 86% of common items showed differences across cohorts. We found administration, scoring, and coding differences for seemingly equivalent items. These differences corresponded to variability across cohorts in score distributions and ranges. We performed data augmentation to adjust for differences.

**CONCLUSION:** Cross-cohort administration, scoring, and procedural differences for cognitive instruments are frequent and need to be assessed to address potential impact on meta-analyses and cognitive data interpretation. Detecting and accounting for these differences is critical for accurate attributions of cognitive health across cohort studies.

**Keywords**

cognition; dementia; epidemiology; methods

## Introduction

Researchers are increasingly performing meta-analyses of individual participant cognitive data from existing large databases and cohorts to better understand cognitive decline and dementia risk [1,2]. Pooling cognitive data from longitudinal population-based cohorts facilitates improved power and novel capabilities to investigate biomedical, lifestyle, and sociocultural factors that may affect cognition across the lifespan [3,4]. Prior to pooling cognitive data, pre-statistical harmonization is necessary to ensure accurate and consistent inferences about cognitive health across studies.

Pre-statistical harmonization [5] of cognitive instruments is a complicated, qualitative process that involves a careful review of cohort characteristics (e.g., subject selection procedures, demographic factors such as race/ethnicity, socioeconomic status, language), and cognitive instruments to identify common and unique items across datasets, and candidate items that might be made comparable with minimal transformation. Specific procedures for pre-statistical harmonization of cognitive instruments across studies are not established [6].

Pre-statistical harmonization of cognitive instruments is a unique challenge because heterogeneity exists in the instruments used to measure cognition, and the procedures for administering them. Over 500 instruments are available for clinical and research use [7], and they differ by domain of assessment, level of measurement precision (e.g., ranging from a 10-minute, in-person or telephone-based cognitive screening assessment to a several-hour battery of neuropsychological tests), sensitivity to change, and other factors. Heterogeneity may also be present across seemingly parallel cognitive instruments, such as differences in test version, test adaptation for individual study purposes, and differences in administration, scoring, and coding procedures. These procedural details are crucial to uncover prior to the harmonization process. Data augmentation procedures are often required to address this variability across studies to create one cohesive meta-analysis dataset. This aspect of pre-statistical harmonization is rarely discussed or documented in harmonization research [6].

To promote reproducible research in cognitive decline and dementia risk, in this report we describe pre-statistical harmonization for meta-analysis of individual participant data from six American population-based cohorts for the Effect of Lower Blood Pressure over the Life Course on Late-life Cognition (BP-COG) study, quantifying the effect of blood pressure (BP) levels over the life course on cognitive trajectories in Black, Hispanic, and White individuals [1,2]. We aim to describe our approach to pre-statistical harmonization of cognitive data, including our detailed review of the administration, scoring procedures, and score ranges of cognitive instruments across cohorts to determine their degree of equivalence and suitability for pooling. Second, we summarize our findings from this detailed review and their implications for data interpretation. Finally, we offer recommendations for

pre-statistical harmonization procedures for cognitive instruments to inform future meta-analyses.

## Methods

### Cohort studies

Six NIH-funded, longitudinal cohort studies were included in the present study: The Atherosclerosis Risk in Communities Study (ARIC [8], the Coronary Artery Risk Development in Young Adults Study (CARDIA;[9]), the Cardiovascular Health Study (CHS; [10]), the Framingham Offspring Study (FOS; [11]), the Multi-Ethnic Study of Atherosclerosis (MESA; [12]), the Northern Manhattan Study (NOMAS; [13]). All participants provided informed consent. A description of each cohort study and the rationale for its inclusion in the larger study is available [1]. We selected these cohorts because all had repeated measures of cognition and BP using similar methods, high-quality data on dementia risk factors, overlapping cognitive instruments, physician-adjudicated dementia diagnosis, availability of brain magnetic resonance imaging, racial, ethnic, and geographic diversity, and included adults aged 18 and older. Four cohorts (NOMAS, ARIC, MESA, CARDIA) enrolled adults in early adulthood to mid-life, which is the time in the life course with the most consistent association between BP and later-life cognitive decline. The other two cohorts enrolled beginning in early life (FOS) and older adulthood (CHS).

### Procedure

First, study neuropsychologists (EB, BG) identified cognitive instruments in each of the six cohort studies from documentation provided by each cohort, available published papers, and scrutiny of cognitive datasets. We considered all cognitive instruments at all study visits for inclusion in the harmonization, with the exception of measures of literacy and premorbid intellectual functioning that are relatively less sensitive to age-related cognitive decline.

After identifying available cognitive instruments, we contacted cohort study investigators to request unpublished administration, scoring, and procedural details of cognitive test batteries. Documentation provided by each cohort study included test forms, data entry forms, and administration and scoring instructions. Procedural details extracted from this process included the published test version, administration and scoring details (e.g., stopping rules; acceptable responses for specific items), possible/theoretical score ranges (based upon the instrument structure and number of items), and metrics available for each instrument (e.g., individual item data, raw and standardized summary scores). We reviewed available raw data for each instrument for score ranges and distributions.

When available and appropriate, we reviewed individual test items comprising cognitive instruments rather than composite/summary test scores (e.g., individual test items for the Mini Mental State Examination [MMSE]; Telephone Interview for Cognitive Status [TICS], Modified Mini Mental State Examination [3MSE], or Montreal Cognitive Assessment [MoCA]). Decomposition of these instruments into their component items enabled us to identify additional common items between cohorts and to link telephone-administered test batteries with in-person test batteries.

Study neuropsychologists reviewed and categorized each cognitive instrument item into relevant cognitive domains (i.e., memory, executive functioning).

### Comparability of items across cohort and language

We identified cognitive test items that were comparable across cohorts based upon detailed documentation. For cohorts that offered test administration in Spanish, neuropsychologists considered the degree of linguistic and cultural equivalence based on the evidence base for individual cognitive instrument items. Items for which there was concern for non-equivalence across English and Spanish were considered separate items by language rather than comparable items.

### Data augmentation

We employed several data augmentation strategies to adjust for differences uncovered during the pre-statistical harmonization process, including alignment of coding procedures, winsorization, and equipercentile equating. The decision of how to address each disparate test item was made on a case-by-case basis by an experienced epidemiologist in consultation with study neuropsychologists, based on the distribution of scores across tests and assuming that we properly identified test items that were comparable across cohorts, except for version or administration differences. To make these decisions, we meticulously scrutinized exploratory analyses of test scores (e.g., dotplots and histograms, stem and leaf plots, tables of minima, maxima, medians, and means) by cohort for every test item presumed to be comparable across cohorts. When procedural or distributional differences across cohorts in common test items were identified, we reviewed each item for possible data augmentation.

## Results

### Summary of cognitive instrument items

We identified 34 cognitive instruments, 13 of which were common across two or more cohorts. Animal naming spanned all six cohorts, letter fluency spanned five cohorts, and four instruments (Mini Mental State Examination, Boston naming, digit span, and digit symbol substitution) spanned four cohorts (Table 1). From these 34 cognitive instruments, we identified 153 items (Supplemental Tables 1 and 2), 64 of which were common across at least two cohorts. A detailed review of instrument administration and scoring procedures across cohorts revealed several sources of procedural differences across cohorts, despite seemingly common tests (Supplemental Tables 1 and 2). Frequent sources of such heterogeneity included differences in instrument version, instrument adaptation, administration procedures, data (i.e., type of score provided for a cognitive instrument), and component items.

### Instrument Version

Differences in instrument version affected 10 of 13 common instruments. For example, four studies (ARIC, CHS, CARDIA, MESA) implemented a speeded task requiring transcription of symbols that corresponded to specific digits, titled Digit Symbol Substitution Test for the Wechsler Adult Intelligence Scale-Revised (WAIS-R, [14]) published in 1981. Two studies (MESA and CARDIA) used an updated version, the Digit Symbol Coding Test, published

in the test's 1997 revision (Wechsler Adult Intelligence Scale- III; WAIS-III, [15]. Although these tests are nearly equivalent in structure, they had different times to complete (90 sec for WAIS–R vs. 120 sec for WAIS–III) and a different number of possible items (93 items for WAIS–R vs. 133 items for WAIS–III). Of note, a similarly-titled, but distinct, measure was administered in the NOMAS study (Symbol Digit Modalities Test; SDMT [16]), which allowed 90 sec to transcribe numbers corresponding to distinct symbols. Scores on this test tend to be lower than for the digit symbol versions, possibly due to subtle differences in task demands[17]. For each of these tests, performance is quantified as the number of correct items completed; as such, an equivalent raw score is associated with a different level of performance (i.e., speed) across these studies. For example, 50 items completed with a 90 sec time limit corresponds to 0.56 items/sec, whereas 50 items completed within a 120 sec time limit corresponds to 0.42 items/sec.

### Instrument adaptation

Adaptations from standard administration or scoring of tests impacted 8 of 13 common instruments. For example, the California Verbal Learning Test (CVLT) was administered in two cohorts (NOMAS, CHS Cognition Study). The CHS Cognition Study used the standard CVLT-1 administration procedures, with a list of 16 words and instructions to recall items on a shopping list. The NOMAS study, in contrast, used a list of 12 rather than 16 words; the words were presented via audio recording with a male or female voice, and respondents were required to recall the words and whether the word was presented in a male or female voice.

### Administration procedures

Differences in administration procedures, such as stopping rules for timed tests, were noted in 6 of 13 common instruments. Some administration differences led to different raw score ranges across common tests. For example, the Trail Making Test, Part B had a maximum raw score of 240 sec in the ARIC study due to their discontinuation rule of 240 sec and/or 5 errors. This same test had a maximum raw score of 572 sec in the CHS Cognition study and 600 sec in the FOS study, for which stopping rules were not available in study documentation. Studies also varied with procedures for the items evaluating working memory (e.g., serial subtractions or spelling backwards) on the MMSE: NOMAS administered serial subtraction only when full credit was not awarded for spelling backwards; CHS studies administered both items and scored the higher of the two responses, whereas ARIC and NOMAS administered only the spelling backwards item.

### Score type

Studies varied in the selection of summary scores provided for common instruments, and most instruments had several summary scores included in datasets. For 1 of 13 common instruments, there were no equivalent summary scores available for common instruments across studies. Specifically, for the digit span test, the FOS study provided the maximum string of digits correctly answered, whereas ARIC and MESA provided the total score for the measure, and CHS provided both metrics.

### Scoring and coding procedures

Review of item scoring and coding procedures revealed several sources of heterogeneity across cohorts and affected items in 2 of 13 common instruments (MMSE and TICS), in addition to common items identified from other instruments (3MSE, CASI, MoCA). Studies varied with regard to the degree of precision required by the respondent to award credit for an item. For example, for the orientation to season item, NOMAS awarded credit if respondents made an error during the first or last day of the month, whereas CHS, FOS, and ARIC required precise responses. Common items were also identified between the MMSE and other cognitive screening instruments (i.e., 3MSE, Cognitive Abilities Screening Instrument (CASI), MoCA). Across these measures, differences with regard to scoring and coding were noted frequently across equivalent items. For example, when orientation to month was administered in the CASI (MESA), coding was as follows: 2= accurate within 5 days; 1= missed by 1 month, and 0 = missed by 2+ months; whereas coding of this item for MMSE and TICS (ARIC, CHS, NOMAS) was 1=correct, 0=incorrect (Supplemental Table 2).

### Cultural and linguistic equivalence

We reviewed candidate comparable items that were administered across English and Spanish. We noted evidence of differential difficulty across language for some test items. For example, items from the Boston Naming Test have shown differential difficulty across English and Spanish [18]. Because of concern for possible non-equivalence in data interpretation across English and Spanish, we considered these test items as separate rather than comparable items.

### Data augmentation

Supplemental Tables 1 and 2 describe the data augmentation strategies used for each item included in the harmonization. Of the 64 comparable items identified across cohorts, 55 required data augmentation. For some instruments, such as the letter fluency test and the Trail Making Test, we transformed scores to curtail outliers via winsorization (i.e., pulling in extreme values) (Supplemental Table 1). We coarsened more finely measured test scores in some cohorts in order to map them to other available test items in other cohorts (e.g., creating summary scores comprised of individual test items); for example, this was done for test items such as delayed word recall for 3MSE from the CHS study (Supplemental Table 2). For tests with completion time scores, such that higher scores are indicative of worse performance (e.g., Trail Making Test, Stroop Color and Word Test, Grooved Pegboard Test), we adjusted the direction of the scores to reflect higher scores indicative of better performance (Supplemental Table 1). We used equipercentile equating [19] to adjust the distributions of like test items across cohorts to be on a common scale for tests such as the CVLT and Digit Symbol Substitution Test/Digit Symbol Coding Test (Supplemental Table 1). To address differences in coding procedures for correct items (e.g., CASI items, MESA cohort), we recoded items to align across studies (Supplemental Table 2). To address differences in age distributions across cohorts, the age range was restricted to 53–80 years to develop the equating algorithm. The equating algorithm was then applied to the full sample.

To illustrate the impact of the sources of heterogeneity on cognitive data and data augmentation procedures, we selected data from the DSST/DSC instrument (Figure 1). As illustrated in Figure 1 (panel B), instrument test version differences were associated with markedly different raw score distributions across cohorts. To address these differences in raw scores (and given equivalence of construct assessed), we used equipercentile equating to align distributions (Figure 1, Panel C).

## Discussion

In the pre-statistical harmonization of cognitive instruments for meta-analyses of individual participant data from six longitudinal cohort studies, we found numerous differences in cognitive instrument version, administration, scoring, coding, and other procedures across cohorts. We found differences even in seemingly comparable cognitive instruments across cohorts that made them potentially nonequivalent. We found procedural differences even for widely used, standardized cognitive tests, such as the MMSE. Many of these sources of cognitive instrument variability were associated with clear differences in cognitive data distributions across cohorts. Of the 64 comparable items identified across cohorts, 55 (86%) required data augmentation to facilitate comparability across cohorts.

Although studies have examined statistical harmonization of cognitive data across studies [20–22], no studies have examined pre-statistical harmonization of cognitive data across studies[6]. Our results provide evidence that careful pre-statistical harmonization, including detecting and accounting for cognitive instrument procedural differences across cohorts, is a critical analytic step of the harmonization process. Assuming equivalence could lead to misattribution of procedural sources of variance to systematic differences between groups of people across cohorts. In addition, transparency in the pre-statistical harmonization process is critical for reproducible research and to ensure scientific scrutiny of this aspect of harmonization methodology.

The present study highlights the critical need for performing specific, detailed and comprehensive pre-statistical harmonization steps prior to statistical harmonization, as we summarized in a checklist (Table 2). These procedures are warranted for all approaches to cognitive data harmonization. At each level of pre-statistical harmonization, expert content reviewers (e.g., neuropsychologists) are needed to ensure that decisions made about the equivalence and utility of test items are appropriate. This expert content review process is particularly necessary prior to harmonization of cognitive instruments across populations that have diverse cultural and linguistic characteristics, as expert reviewers may identify cultural and linguistic sources of non-equivalence across items that could lead to misattributions of cognitive test score differences if not detected.

Our findings have important implications for future studies that aim to harmonize data from cognitive instruments. We summarize our recommended procedures for pre-statistical harmonization of cognitive instruments in Table 2 and Figure 2. We outline pre-statistical harmonization steps, critical information to document, relevant examples from the BP COG study, and offer possible solutions when differences are uncovered. We offer the following conclusions and recommendations for future cognitive data

harmonization studies based upon this work. First, use of parallel, standardized cognitive tests does not guarantee identical implementation of these tests across studies [23]; we recommend careful scrutiny and documentation of procedural differences prior to data pooling. This process is resource-intensive, should involve a content expert in neuropsychological assessment, and should be budgeted for when planning studies requiring cognitive instrument harmonization. Availability of source documents pertaining to cognitive instrument administration and scoring procedures are critical for comprehensive and thorough pre-statistical harmonization. We only know what we know: undocumented deviations in standard administration are difficult to detect post-hoc for investigators who were not involved during original data collection. Second, heterogeneity in implementation of common cognitive instruments can lead to important differences in score distributions across studies. Failure to detect and account for these differences during pre-statistical harmonization could lead to erroneous attributions regarding cognitive health across cohort studies. Finally, detecting these differences facilitates use of statistical procedures, in IRT-based models, to account for these differences and allow for equating scores.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This work was supported by the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Department of Health and Human Service (R01 NS102715). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health. Representatives of the funding agency have been involved in the review of the manuscript but not directly involved in the collection, management, analysis or interpretation of the data. Additional funding was provided by National Institute of Aging (NIA) grant R01 AG051827 (Levine), NIA Claude Pepper Center grant P30 AG024824 (Galecki, Kabeto), NIA grant K01 AG050699 (Gross), and NIA Michigan Alzheimer's Disease Core Center grant P30 AG053760 (Giordani).

Cohort Funding/Support: The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, HHSN268201700004I). Neurocognitive data is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. The authors thank the staff and participants of the ARIC study for their important contributions.

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). The CARDIA cognitive ancillary study was supported by NIA R01 AG063887. This manuscript has been reviewed by CARDIA for scientific content.

The Cardiovascular Health Study (CHS) was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, 379 N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, 380 N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

The Framingham Heart Study is a project of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This project has been funded in whole or in part with Federal

funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under contract No. HHSN268201500001I.

The Northern Manhattan Study is funded by the National Institutes of Health, National Institute of Neurological Disorders and Stroke (R01 NS29993).

Multi-Ethnic Study of Atherosclerosis (MESA) was supported by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by grants UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 from the National Center for Advancing Translational Sciences (NCATS). The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

Conflicts of Interest/Disclosure Statement

Dr. Briceño reports a grant from the National Institute on Aging during the conduct of the study. Dr. Levine reports grants from the National Institutes of Health (NIH) during the conduct of the study and outside the submitted work. Ms. Hingtgen report grants from the NIH/National Institute of Neurological Disorders and Stroke (NINDS) during the conduct of the study. Dr. Manly reports grants from NIH/National Institute on Aging during the conduct of the study. Dr. Gottesman reports other support from the American Academy of Neurology outside the submitted work. Dr Sidney reports a contract from the National Heart, Lung, and Blood Institute outside the submitted work. Dr Yaffe reports grants from the National Heart, Lung, and Blood Institute during the conduct of the study. Dr Sacco reports grants from NINDS during the conduct of the study and grants from Boehringer Ingelheim, American Heart Association, and Florida Department of Health outside the submitted work. Dr. Wright reports grants from NIH/NINDS during the conduct of the study and royalties from UpToDate for 2 chapters on vascular dementia. No other disclosures were reported.

## References

[1]. Levine DA, Gross AL, Briceño EM, Tilton N, Kabeto MU, Hingtgen SM, Giordani BJ, Sussman JB, Hayward RA, Burke JF, Elkind MSV, Manly JJ, Moran AE, Kulick ER, Gottesman RF, Walker KA, Yano Y, Gaskin DJ, Sidney S, Yaffe K, Sacco RL, Wright CB, Roger VL, Allen NB, Galecki AT (2020) Association between Blood Pressure and Later-Life Cognition among Black and White Individuals. JAMA Neurol. 77, 810–819. [PubMed: 32282019]

[2]. Levine DA, Gross AL, Briceño EM, Tilton N, Giordani BJ, Sussman JB, Hayward RA, Burke JF, Hingtgen SM, Elkind MSV, Manly JJ, Gottesman RF, Walker KA, Yano Y, Gaskin DJ, Sidney SS, Sacco RL, Tom SE, Wright CB, Yaffe K, Galecki AT (2021) Sex Differences in Cognitive Decline among US Adults. JAMA Netw. Open 4, e210169. [PubMed: 33630089]

[3]. Hofer SM, Piccinin AM (2010) Toward an Integrative Science of Life-Span Development and Aging. Journals Gerontol. Ser. B Psychol. Sci. Soc. Sci. 65B, 269–278.

[4]. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W (2009) Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychol. Methods 14, 126–149. [PubMed: 19485625]

[5]. Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, Belleville S (2013) Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. In Methods Research Report. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290–2007-10060-I.) AHRQ Publication No.13-EHC040-EF. Rockville, MD: Agency for Healthcare Research and Quality.

[6]. Griffith LE, Van Den Heuvel E, Fortier I, Sohel N, Hofer SM, Payette H, Wolfson C, Belleville S, Kenny M, Doiron D, Raina P (2015) Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. J. Clin. Epidemiol. 68, 154–162. [PubMed: 25497980]

[7]. Lezak MD, Howieson DB, Bigler ED, Tranel D (2012) Neuropsychological assessment, Oxford University Press, New York, NY.

[8]. The ARIC Investigators (1989) The atherosclerosis risk in communities (ARIC) study: Design and objectives. Am. J. Epidemiol. 129, 687–702. [PubMed: 2646917]

[9]. Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR, Liu K, Savage PJ (1988) CARDIA: study design, recruitment, and some characteristics of the examined subjects. J. Clin. Epidemiol. 41, 1105–1116. [PubMed: 3204420]

[10]. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary DH, Psaty B, Rautaharju P, Tracy RP, Weiler PG (1991) The Cardiovascular Health Study: Design and rationale. Ann. Epidemiol. 1, 263–276. [PubMed: 1669507]

[11]. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP (1975) The Framingham offspring study. Design and preliminary data. Prev. Med. (Baltim). 4, 518–525.

[12]. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacobs DR, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP (2002) Multi-Ethnic Study of Atherosclerosis: Objectives and design. Am. J. Epidemiol. 156, 871–881. [PubMed: 12397006]

[13]. Sacco RL, Boden-Albala B, Gan R, Chen X, Kargman DE, Shea S, Paik MC, Hauser WA (1998) Stroke incidence among white, black, and Hispanic residents of an urban community: The Northern Manhattan Stroke Study. Am. J. Epidemiol. 147, 259–268. [PubMed: 9482500]

[14]. Wechsler D (1981) Wechsler Adult Intelligence Scale-Revised, Psychological Corporation, New York, NY.

[15]. Wechsler D (1997) Wechsler Adult Intelligence Scale-Third Edition, The Psychological Corporation, San Antonio, TX.

[16]. Smith A (1982) Symbol Digit Modalities Test, Western Psychological Services, Los Angeles, CA.

[17]. Strauss E, Sherman E, Spreen O (2006) A compendium of neuropsychological tests: Administration, norms, and commentary, Oxford University Press, New York, NY.

[18]. Jahn DR, Mauer CB, Menon CV, Edwards ML, Dressel JA, Obryant SE (2013) A brief Spanish-English equivalent version of the Boston Naming Test: A Project FRONTIER Study. J. Clin. Exp. Neuropsychol. 35, 835–845. [PubMed: 23998641]

[19]. Kolen MJ, Brennan RL (2014) Test Equating, Scaling, and Linking: Methods and Practices, Springer, New York.

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

[21]. 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, 144–153. [PubMed: 24481241]

[22]. Gross AL, Sherva R, Mukherjee S, Newhouse S, Kauwe JSK, Munsie LM, Waterston LB, Bennett DA, Jones RN, Green RC, Crane PK (2014) Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. Neuroepidemiology 43, 194–205. [PubMed: 25402421]

[23]. Gross AL, Inouye SK, Rebok GW, Brandt J, Crane PK, Parisi JM, Tommet D, Bandeen-Roche K, Carlson MC, Jones RN (2012) Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. J. Clin. Exp. Neuropsychol. 34, 758–772. [PubMed: 22540849]

A.

| Cohort Study | Test Version | Test Name | Time Limit | Possible Score Range | Raw Score Range | Equated Score Range |
|---|---|---|---|---|---|---|
| ARIC | WAIS-R | Digit Symbol Substitution Test | 90 sec | 0-92 | 0-92 | 0-93 |
| CARDIA | WAIS-III | Digit Symbol Coding Test | 120 sec | 0-133 | 0-125 | 0-93.9 |
| CHS | WAIS-R | Digit Symbol Substitution Test | 90 sec | 0-92 | 0-90 | 0-90 |
| MESA | WAIS-III | Digit Symbol Coding Test | 120 sec | 0-133 | 0-120 | 0-70.9 |
| NOMAS | n/a | Symbol Digit Modalities Test | 90 sec | 0-110 | 0-67 | 5-64.7 |

D.

| Cohort Study | Test Version | Test Name | Number of items | Possible Score Range | Raw Score Range | Equated Score Range |
|---|---|---|---|---|---|---|
| CHS | WAIS-R (modified) | Block Design | 5 | 0-26 | 0-24 | 0-24 |
| FOS | WAIS (modified) | Block Design | 6 | 0-32 | 0-32 | 0-15.75 |

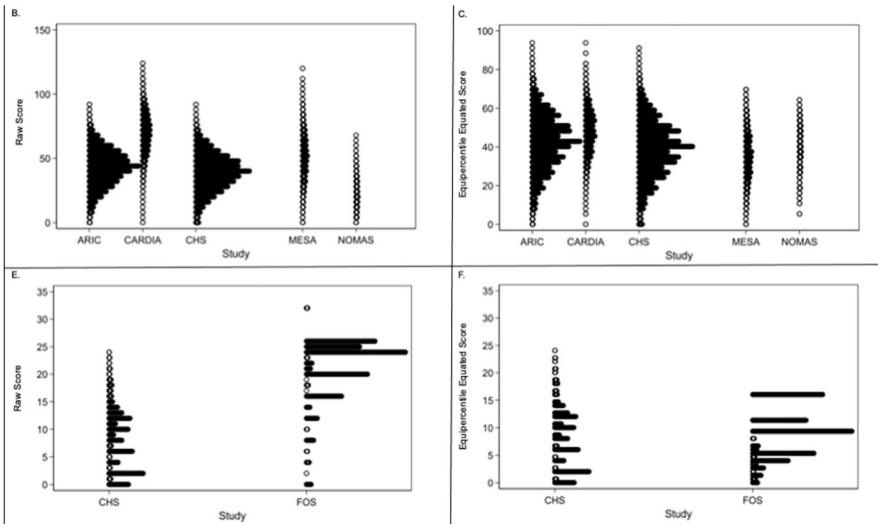


**Figure 1.**

Impact of procedural differences on digit symbol substitution test data across cohorts. Panel A summarizes differences in test version, procedure, and raw score ranges across cohorts for the digit symbol substitution test. Panel B displays the raw score distributions across cohorts for the digit symbol substitution test. Panel C displays the equipercentile-equated distributions for the digit symbol substitution test. All scores are scaled on a T score metric (mean = 50, standard deviation = 10).

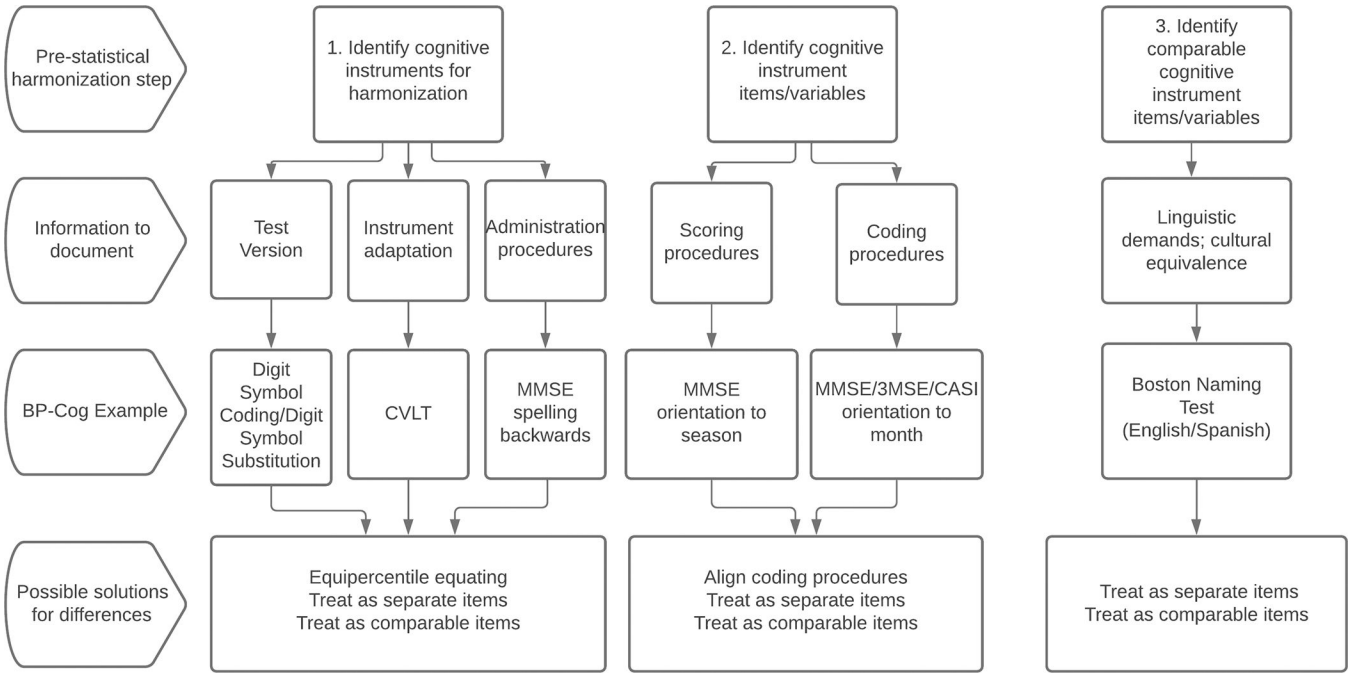


**Figure 2.**

Summary of recommended pre-statistical steps for harmonization of cognitive instruments. Figure outlines recommended pre-statistical procedures for cognitive instrument harmonization, examples of sources of heterogeneity across cohorts, and possible solutions for addressing sources of heterogeneity.

**Table 1.**

Common and unique cognitive instruments administered in each cohort study

| | Domain | ARIC | CARDIA | CHS | FOS | MESA | NOMAS |
|---|---|---|---|---|---|---|---|
| **Global cognitive performance (GCP)** | **Executive function**/ processing speed | DSST | DSC | DSST | | DSC | |
| | | | | | | | SDMT |
| | | Animal Naming | Animal Naming | Animal Naming | Animal Naming | Animal Naming | Animal Naming |
| | | | | Semantic Gen. | | | |
| | | Letter Fluency | Letter Fluency | Letter Fluency | Letter Fluency | | Letter Fluency |
| | | Trail Making | | Trail Making | Trail Making | | |
| | | | Stroop Test | Stroop NST | | | |
| | | | | | | | Color Trails; Odd Man Out |
| | | | | Baddeley-Papagno | | | |
| | | | | | WAIS Similarities | | |
| | | WAIS-R Digit Span | | WAIS-R Digit Span | WAIS Digit Span | WAIS-III Digit Span | |
| | | | | | | | Digit ordering; WAIS-III LNS |
| | **Learning/Memory** | | RAVLT | CVLT | | | CVLT-II |
| | | DWR | | | | | |
| | | | | | WMS Paired Assoc. | | |
| | | WMS-R LM | | | WMS LM | | |
| | | | | Rey O CFT Recall | | | |
| | | | | | WMS-R VR | | |
| | General Mental Status | MMSE | | MMSE | MMSE | | MMSE |
| | | | MoCA | | | | |
| | | TICS | | TICS | | | TICS |
| | | | | 3MSE | | | |
| | | | | | | CASI | |
| | Language | Boston Naming (30-item) | | Boston Naming (30-item) | Boston Naming (30-item) | | Boston Naming (15-item) |
| | Motor | | | Grooved Pegboard | | | Grooved Pegboard |
| | | | | | Finger Tapping | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Visuospatial | Clock Time Perception | | | | | |
| | | | | WAIS-R Block Design | WAIS Block Design | | |
| | | | | Ravens | | | |
| | | | | Rey O CFT Copy | | | |
| | | | | | Hooper VOT | | |
| | | | | | | | VMI |

*Note*: Cognitive instruments were administered to different participant samples and at different time intervals across the 6 cohorts. Instruments identified as candidate comparable items are listed in the same row.

Abbreviations:

Cohorts: CARDIA is the Coronary Artery Risk Development in Young Adults study. NOMAS is the Northern Manhattan Study. ARIC is the Atherosclerosis Risk in Communities Study. CHS is the Cardiovascular Health Study. MESA is the Multi-Ethnic Study of Atherosclerosis study. FOS is the Framingham Offspring Study.

Cognitive Instruments: 3MSE is the Modified Mini Mental State Examination. ANT is Animal Naming Test. BNT is Boston Naming Test. CASI is Cognitive Abilities Screening Instrument. CVLT is California Verbal Learning Test. DSC is WAIS-III Digit Symbol Coding Test. DSST is the WAIS-R Digit Symbol Substitution Test. DWR is the Delayed Word Recall test. Hooper VOT is Hooper Visual Organization Test. LFT is Letter Fluency Test. LM is Logical Memory. LNS is Letter-Number Sequencing. MMSE is Mini-Mental State Examination. MoCA is Montreal Cognitive Assessment. RAVLT is Rey Auditory Verbal Learning Test (RAVLT). Rey O CFT is Rey-Osterrieth Complex Figure Test. SDMT is Symbol Digit Modalities Test. Semantic Gen. is semantic word generation (birds, dogs). Similarities is the WAIS Similarities test. Stroop NST is the Stroop Neuropsychological Screening Test. TICS is the Telephone Interview for Cognitive Status. VMI is Developmental Test of Visuomotor Integration. VR is Visual Reproduction. WAIS is Wechsler Adult Intelligence Scale. WAIS-R is Wechsler Adult Intelligence Scale-Revised. WAIS-III is Wechsler Adult Intelligence Scale-3rd Edition. WMS is Wechsler Memory Scale. WMS-R is WMS-Revised. WMS-III is WMS-3rd Edition.

**Table 2.**

Checklist of procedures for pre-statistical harmonization of cognitive data

| |
|---|
| Define and state the study question |
| Identify cohorts based on the study question |
| Identify cognitive instruments within each cohort |
| Collect all available documentation about cognitive instruments |
| Test versions |
| Test adaptations |
| Administration procedures |
| Stopping rules for timed tests |
| Items administered conditional on other items |
| Scoring and coding procedures |
| Theoretical and empirical score ranges |
| Language of administration |
| Cultural and linguistic comparability |
| Mode of administration (phone vs in-person vs internet) |
| Consider test items from cognitive instruments, rather than summary scores, when available and reasonable |
| Expert neuropsychologists to categorize cognitive test items into domains of interest, informed by the study question |
| Identify items common across cohorts based on available documentation |
| Consider data transformation strategies to derive common items across cohorts if items are conceptually comparable but different in distribution or with different response values |
| Exploratory data analysis |
| Winsorization |
| equipercentile equating |