Neuropsychology

Cognitive Domain Harmonization and Cocalibration in Studies of Older Adults

Shubhabrata Mukherjee, Seo-Eun Choi, Michael L. Lee, Phoebe Scollard, Emily H. Trittschuh, Jesse Mez, Andrew J. Saykin, Laura E. Gibbons, R. Elizabeth Sanders, Andrew F. Zaman, Merilee A. Teylan, Walter A. Kukull, Lisa L. Barnes, David A. Bennett, Andrea Z. Lacroix, Eric B. Larson, Michael Cuccaro, Shannon Mercado, Logan Dumitrescu, Timothy J. Hohman, and Paul K. Crane

Online First Publication, August 4, 2022. http://dx.doi.org/10.1037/neu0000835

AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

ISSN: 0894-4105

https://doi.org/10.1037/neu0000835

Cognitive Domain Harmonization and Cocalibration in Studies of Older Adults

Shubhabrata Mukherjee[1], Seo-Eun Choi[1], Michael L. Lee[1], Phoebe Scollard[1], Emily H. Trittschuh[2, 3], Jesse Mez[4], Andrew J. Saykin[5], Laura E. Gibbons[1], R. Elizabeth Sanders[1], Andrew F. Zaman[6], Merilee A. Teylan[7], Walter A. Kukull[7, 8], Lisa L. Barnes[9], David A. Bennett[9], Andrea Z. Lacroix[10], Eric B. Larson[11], Michael Cuccaro[6], Shannon Mercado[12, 13], Logan Dumitrescu[12, 13], Timothy J. Hohman[12, 13], and Paul K. Crane[1]

[1] Department of Medicine, The University of Washington
[2] Department of Psychiatry and Behavioral Sciences, The University of Washington
[3] VA Puget Sound Health Care System, Seattle, Washington, United States
[4] Department of Neurology, Boston University School of Medicine
[5] Department of Radiology and Imaging Services, Indiana Alzheimer's Disease Research Center, Indiana University
[6] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine
[7] National Alzheimer's Coordinating Center, Department of Epidemiology, University of Washington
[8] Department of Epidemiology, The University of Washington
[9] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, United States
[10] Department of Epidemiology, University of California San Diego
[11] Kaiser Permanente Washington Health Research Institute, Seattle, Washington, United States
[12] Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Nashville, Tennessee, United States
[13] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States

on behalf of Adult Changes in Thought (ACT), Alzheimer's Disease Neuroimaging Initiative (ADNI), Religious Orders Study (ROS), Memory and Aging Project (MAP), Minority Aging Research Study (MARS), National Alzheimer's Coordinating Center (NACC)

Objective: Studies use different instruments to measure cognitirating cognitive tests permit direct comparisons of individuals across studies and pooling data for joint analyses. ***Method:*** We began our legacy item bank with data from the Adult Changes in Thought study ($n = 5{,}546$), the Alzheimer's Disease Neuroimaging Initiative ($n = 3{,}016$), the Rush Memory and Aging Project ($n = 2{,}163$), and the Religious on such as the Mini-Mental State Examination, the Alzheimer's Disease Assessment Scale–Cognitive Subscale, the Wechsler Memory Scale, and the Boston Naming Test. CocalibOrders Study ($n = 1{,}456$). Our workflow begins with categorizing items administered in each study as indicators of memory, executive functioning, language, visuospatial functioning, or none of these domains. We use confirmatory factor analysis models with data from the most recent visit on the pooled sample across these four studies for cocalibration and derive item parameters for all items. Using these item parameters, we then estimate factor scores along with corresponding standard errors for each domain for each study. We added additional studies to our pipeline as available and focused on thorough consideration of candidate anchor items with identical content and administration methods across studies. ***Results:*** Prestatistical harmonization steps such qualitative and quantitative assessment of granular cognitive items and evaluating factor structure are important steps when trying to cocalibrate cognitive scores across studies. We have cocalibrated cognitive data and derived scores for four domains for 76,723 individuals across 10 studies. ***Conclusions:*** We have implemented a large-scale effort to harmonize and cocalibrate cognitive domain scores across multiple studies of cognitive aging. Scores on the same metric

Shubhabrata Mukherjee https://orcid.org/0000-0003-2522-2884

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data analyses were funded by U24 AG074855 (Timothy J. Hohman, Michael Cuccaro, A. Toga, MPI), U01 AG068057 (P. Thompson, C. Davatzikos, H. Huang, Andrew J. Saykin, L. Shen, MPI), and R01 AG057716 (Timothy J. Hohman, PI). Paul K. Crane, Eric B. Larson, and Andrea Z. Lacroix were funded by U01 AG0006781 and U19 AG066567 (Eric B. Larson, Paul K. Crane, Andrea Z. Lacroix, MPI). Paul K. Crane was also supported by R01 AG029672 (Paul K. Crane, PI). Shubhabrata Mukherjee was also supported by K25 AG055620. Jesse Mez's efforts were also supported by R01 AG061028 (Jesse Mez, K. Dams-O'Connor, MPI). Laura E. Gibbons' efforts were also supported by P30 AG066509 (T. Grabowski, PI). Andrew J. Saykin's efforts were also supported by U01 AG042904 (M. Weiner, PI), P30 AG010133 (Andrew J. Saykin, PI), and R01 AG019771 (Andrew J. Saykin, PI). Logan Dumitrescu was supported by R01 AG073439. ACT data collection was supported by U19 AG066567 (Eric B. Larson, Paul K. Crane, Andrea Z. Lacroix, MPIs). ADNI data collection and genotyping were supported by U01 AG024904 (M. Weiner, PI). Data

continued

facilitate meta-analyses of cognitive outcomes across studies or the joint analysis of individual data across studies. Our systematic approach allows for cocalibration of additional studies as they become available and our growing item bank enables robust investigation of cognition in the context of aging and dementia.

Key Points

Question: What considerations were addressed in setting up and implementing a robust workflow that harmonizes and cocalibrates cognitive data across studies of older adults? ***Findings:*** Data from thousands of individuals at tens of thousands of study visits have been cocalibrated to the same metrics for four different cognitive domains. ***Importance:*** These data will facilitate analyses of cognition across studies, despite varying levels of overlap in cognitive tests used across studies. ***Next Steps:*** Cocalibrated scores and standard errors for the studies incorporated in our item banking efforts to date are available to investigators. Additional studies will be incorporated in the coming years using the same methods.

Keywords: cognition, psychometrics, cocalibration, aging, neuropsychological testing

Supplemental materials: https://doi.org/10.1037/neu0000835.supp

Many studies of older adults include tests of cognitive function that are administered to study participants at study visits. Neuropsychological tests used to measure cognition vary across studies (see Supplemental Tables 1–19; Bennett et al., 2018; Montine et al., 2012; Weiner et al., 2017), which make harmonization of data across studies a particular challenge.

Harmonization describes a process of addressing differences in measurement or assessment that could involve procedural, rational, or statistical approaches (Gatz et al., 2015; Gross et al., 2018). Modern psychometric approaches (Borsboom, 2005; Embretson & Reise, 2000; McDonald, 1999) can be used to harmonize cognitive data from different studies. These tools have many desirable features we will illustrate in this article. At the end of our workflow, these tools enable us to derive cocalibrated scores for each cognitive domain. To perform statistical harmonization of cognitive items, we used cocalibration based on confirmatory factor analysis.

collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award Number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. ROS data collection and genotyping were supported by P30 AG10161 (David A. Bennett, PI) and R01 AG15819 (David A. Bennett, PI). MAP data collection and genotyping were supported by R01 AG17917 (David A. Bennett, PI). MARS data collection was supported by RF1 AG22018 (Lisa L. Barnes, PI). NACC data collection was supported by U24 AG072122 (Walter A. Kukull, PI).

Shubhabrata Mukherjee played lead role in conceptualization, formal analysis, investigation, methodology, software, validation and visualization and equal role in project administration, supervision, writing of original draft and writing of review and editing. Seo-Eun Choi played supporting role in formal analysis, investigation, methodology, validation and writing of review and editing and equal role in software. Michael L. Lee played supporting role in formal analysis, investigation, methodology, software and writing of review and editing. Phoebe Scollard played supporting role in formal analysis, investigation, methodology, software and writing of review and editing. Emily H. Trittschuh played lead role in conceptualization, supporting role in writing of review and editing and equal role in investigation. Jesse Mez played lead role in conceptualization, supporting role in writing of review and editing and equal role in investigation. Andrew J. Saykin played lead role in funding acquisition, supporting role in conceptualization and writing of review and editing and equal role in investigation. Laura E. Gibbons played supporting role in investigation and writing of review and editing and equal role in conceptualization, methodology, software, supervision and validation. R. Elizabeth Sanders played lead role in data curation and supporting role in project administration, visualization and writing of review and editing. Andrew F. Zaman played supporting role in writing of review and editing. Merilee A. Teylan played supporting role in data curation. Walter A. Kukull played supporting role in funding acquisition and writing of review and editing. Lisa L. Barnes played supporting role in funding acquisition and writing of review and editing. David A. Bennett played supporting role in funding acquisition and writing of review and editing. Andrea Z. Lacroix played supporting role in funding acquisition and writing of review and editing. Eric B. Larson played supporting role in funding acquisition and writing of review and editing. Michael Cuccaro played lead role in funding acquisition and supporting role in writing of review and editing. Shannon Mercado played supporting role in project administration and writing of review and editing. Logan Dumitrescu played supporting role in funding acquisition, resources and writing of review and editing. Timothy J. Hohman played lead role in funding acquisition and supporting role in resources and writing of review and editing. Paul K. Crane played lead role in conceptualization, funding acquisition, investigation, methodology and project administration and equal role in supervision, validation, writing of original draft and writing of review and editing.

Correspondence concerning this article should be addressed to Shubhabrata Mukherjee, Department of Medicine, The University of Washington, Box 359780, 325 Ninth Avenue, Seattle, WA 98104, United States. Email: smukherj@uw.edu

Cocalibration means "calibrated together." Cocalibrating items in an item bank enables us to obtain scores that are on the same metric, regardless of whether there was total overlapping content in all the specific items administered. Cocalibration should be understood as a particular type of harmonization. Cocalibration facilitates either meta-analysis or pooled analyses of individual-level data.

These psychometric approaches have been used in high stakes educational testing since the 1960s (Lord & Novick, 1968). This same item banking approach enables test forms with no overlapping content to be administered to students who sit across from each other at a testing center. The scores those students receive from their responses are on the same metric, even though they each responded to a distinct set of items. Those items had previously been cocalibrated with each other and many other items in an item bank (Hambleton et al., 1991).

While item banking strategies are an appealing approach to the challenges we faced, there are important differences between educational testing and cognitive testing of older adults. Multiple choice response options are common in educational testing settings and essentially never used in cognitive testing in older adults. Instead in cognitive tests, a wide variety of response formats have been developed (Gruhl et al., 2013), including counts of successful responses in a particular time, time to completion of a task, and scores based on the number of elements of a complex figure that are correctly copied or recalled, to name just a few. Furthermore, in many cases a common stimulus is used in multiple trials, which will lead to correlated item response data beyond the correlation due to a relationship with an underlying domain tested by the trials. This residual correlation can be called a methods effect. For example, trials of a word list learning task will have scores that are more correlated with each other than the correlation of any of those learning trials with any other test of memory. Another example of a methods effect is multiple tasks with very similar formats, such as testing letter fluency with the letters F, A, and S. Scores from those three stimuli will be more closely correlated with each other than any of them with some other measure of language because of the commonality of the tasks.

Our group and others over the past decades have adapted more flexible response formats into our models (Gruhl et al., 2013) and have made extensive use of bifactor models to address secondary domain structures induced by methods effects (R. D. Gibbons et al., 2007). Educational testing also faces an analogous challenge. Many tests of reading comprehension use a single block of text with several items addressing that block of text. Scores from those items are more closely correlated with each other than they are with any other item from the test due to what is known as a "testlet" design. The approaches we have taken to address secondary domain structure induced by methods effects are directly analogous to those used in educational testing settings to address testlets (Li et al., 2006; Wainer et al., 2007).

In this protocol article, we describe the workflows we established for item banking of cognitive test items across studies of older adults. In particular, we use this as an opportunity to discuss the rationale for the choices we have made with greater depth than we have had the opportunity to do previously (Mukherjee et al., 2020). We also report on recent progress integrating data from even more studies of older adults. Taken together, these steps have already facilitated analyses that would not be possible without our efforts at cocalibration. We hope they will be widely used by Alzheimer's researchers in the coming years.

Method

Overview. We have divided our workflow into distinct steps, as summarized in Figure 1. We will discuss each of the steps in the figure sections.

Preliminary Analyses in Each Data Set Considered Separately

The goals of the initial steps in the workflow are to ensure that we have a good understanding of the data, that we have made any transformations to the data needed to integrate the data into our workflows, and that the new data are consistent with our overarching modeling strategy.

Step A1: Acquire Data and Documentation From Each Study

We establish data use agreements for each study and acquire granular level data from cognitive batteries along with detailed documentation on each of the items in the battery. Information that has proven to be useful includes versions of tests, specific stimuli administered, and information on how responses are coded. We mine information from data dictionaries and cognitive administration protocol documents from the parent studies to help us in this process. This step in many cases takes multiple iterations as we learn more about the data set.

Step A2: Domain Assignment

We began our item bank by combining data from four very large studies—the Adult Changes in Thought (ACT) study (n = 5,546), the Alzheimer's Disease Neuroimaging Initiative (ADNI; n = 3,016), the Religious Orders Study (ROS; n = 1,456), and the Rush Memory and Aging Project (MAP; n = 2,163). We refer to this group of studies as "legacy studies."

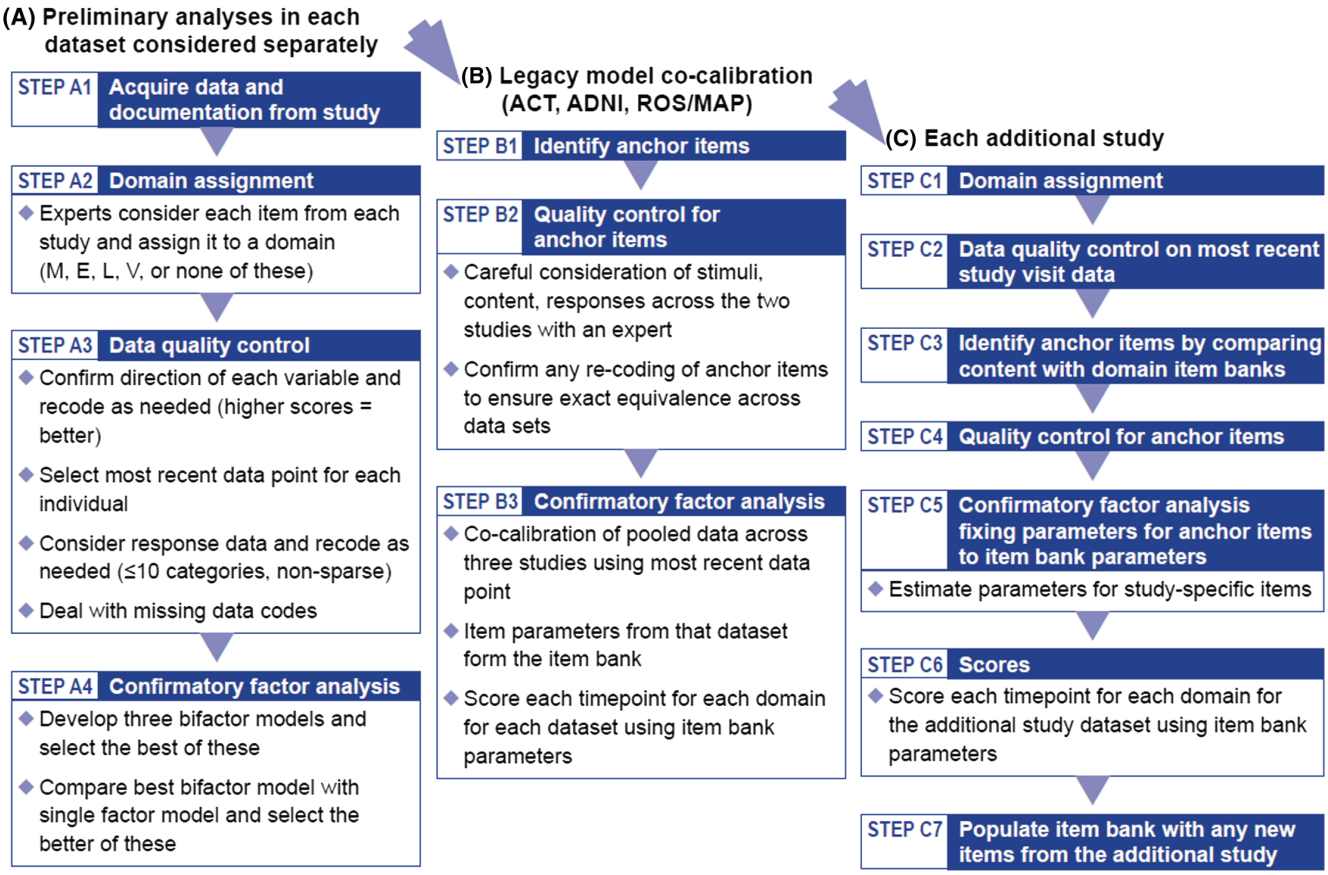
In each of the legacy studies, the expert panel (ET, JM, and AS) assigned items from the cognitive battery to one of the following domains: memory, language, executive functioning, or visuospatial functioning. The studies administer many other items to participants as well, including assessments of subjective impairment or premorbid abilities. We identified items assessing these other domains as well but did not consider them further in our item banking efforts.

If applicable, the expert panel also assigned each of the cognitive items to subdomains based on the cognitive processes involved in each task. For example, the Rey Auditory Verbal Learning Test (RAVLT) trial items were identified as representing the memory domain and the subdomain of verbal episodic encoding while the RAVLT delayed recall item is in memory domain and the verbal episodic retrieval subdomain.

Using study operational and administration manuals, as well as published results, we made note of differing versions and administration methods, so as to be very clear which specific items were administered to study participants at each time point.

Several neuropsychological tests administer items across multiple domains to assess global cognition; examples of this include the Mini-Mental State Examination (MMSE; Folstein et al., 1975), the Modified MMSE (3MS; Teng & Chui, 1987), the Cognitive Abilities Screening Instrument (CASI; Teng et al., 1994), the Community

Figure 1
Cocalibration Workflow



Note. Each of these steps is explained in more detail below. M = memory; E = executive functioning; L = language; V = visuospatial functioning; ACT = Adult Changes in Thought; ADNI = Alzheimer's Disease Neuroimaging Initiative; ROS/MAP = Religious Orders Study/Memory and Aging Project.

Screening Interview for Dementia (CSI-"D"; K. S. Hall et al., 1993; K. S. Hall et al., 2000), and the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005). Many of these scales have overlapping content, and we previously cocalibrated them using similar approaches to those described here (Crane et al., 2008). We took a different approach to these global tests in our current item banking procedures in that we considered each cognitive domain separately. Test items assessing memory would be considered with other memory items and disaggregated from items assessing any other domain.

There are several reasons our thinking on global cognition has evolved since our earlier work (Crane et al., 2008). Conceptually, there are important contrasts across cognitive domains in older adults, and particularly in people with dementia and Alzheimer's disease. A global score necessarily glosses over any distinctions across domains, which may limit understanding of associations with particular brain processes. Consideration of the designs of these tests and the stimuli they use also leads us to derive separate scores for cognitive domains rather than attempting to summarize overall cognition with a single number. Our previous article discussed the interlocking pentagons item and its different treatment in different tests (Crane et al., 2008). It is scored as correct versus incorrect, one point versus zero points, in the MMSE and the CSI "D". The MMSE total score is 30, so the one point for the pentagons (copy two interlocking pentagons) item corresponds with 3.3% of the total score. In the CASI and the 3MS, the same pentagons item is scored on a 0–10 scale, and the total scores go to 100 points, so the pentagons item reflects 10% of the total. The pentagons item is the only element tapping visuospatial functioning in the MMSE, CASI, and 3MS. Should visuospatial functioning represent 3.3% of global cognition? Or 10%? And why? Articles describing the development of these tests (Folstein et al., 1975; Teng & Chui, 1987; Teng et al., 1994) do not provide compelling rationales for the relative importance of each item.

For these reasons, in recent years we moved toward breaking down tests of global cognition into component parts and considering each cognitive domain separately. An investigator wishing to study global cognition who wanted to weight it as four parts memory, two parts language, two parts executive functioning, and one part visuospatial could use the scores we produce to generate such global composite scores. Thus, our approach does not preclude the possibility of studying global cognition and, in fact it enables lines of research that are not possible with global composite scores alone.

Our approach began with theory as directed by our panel of a behavioral neurologist and two neuropsychologists. All three panel members have extensive experience in the clinical and research

evaluation of older adults and neurodegenerative conditions, and their domain assignments reflected this disciplinary background and clinical and research experience. Our goal was to ascertain whether the data available to us from studies of older adults were consistent with this theory, and then, if so, to obtain cocalibrated nonoverlapping domain scores to facilitate cross-study analyses. Others with different goals could have used the data in different ways, such as beginning with exploratory factor analysis approaches or permitting items to load on multiple domains.

Step A3: Data Quality Control

Following domain assignment, our data manager (RES) performed initial quality control steps on the data which included making a master file for all cognitive data with labels and descriptions for each item in the data set. Follow-up quality control steps performed by the analysts (S-EC, ML, PS, SM) included recoding of the data where necessary. For example, items such as Trail Making Tests A and B were reverse coded (i.e., where a higher value indicated worse performance). We checked each item to make sure higher values represent better cognitive performance, and reverse coded as needed. This step facilitates interpretation of factor loadings, as all loadings on the general factor should be positive if all of the items are coded in the same direction, and a negative loading would indicate a need for extra scrutiny.

Whenever possible, we used granular data from each study as it is more informative than summary totals. For example, for a word list learning measure, one can imagine multiple ways of recording participant responses. Ideally, studies could report whether each specific word was recalled on each trial. However, this granular level data of participant response is not always available. Many studies report only the total number of words recalled on each trial or the total of words recalled across all of the learning trials. These scores may be impossible to reconcile across studies unless they were obtained precisely the same way in the two studies. Sometimes data are not electronically available in a sufficiently granular form, in which case we seek resources for data entry or, if that proves impossible, we may decide to drop the item from further cocalibration steps. Collection and data entry of granular data up front helps us derive cognitive scores which are more precise and enable us to be confident in confirming that an item can be used as an anchor, as we will discuss.

Given that we were working with longitudinal data, we had to decide which visit (e.g., first visit, most recent visit) we would use in cocalibration. We selected the most recent visit for each participant. This choice optimizes the spread of cognitive abilities in the data set, which is desirable for ensuring parameters are valid over the entire range of ability levels, while preserving sample size (Embretson & Reise, 2000; Hambleton et al., 1991) and still including only a single observation per person. Some studies such as ACT enrolled people known to be free of dementia, and others enrolled people with particular diagnoses who met with specific eligibility criteria (e.g., ADNI and others). By choosing the most recent visit, cognition in some participants would have declined to the maximal extent available in these data, optimizing the spread of ability levels.

We considered the distribution of each item among participants with nonmissing data and combined categories as needed. Our goals for combining categories were (a) to avoid sparse categories, which we operationally defined as <5 responses per category for each study administering each item, (b) to have no more than 10 categories, the maximum number of categories handled by Mplus v7.4 (Muthen & Muthen, 1998–2012), and (c) to retain the full range of responses from each study, to avoid collapsing categories at the highest and lowest levels of functioning. Retaining variability at the tails at the expense of the center of the distribution minimizes potential floor and ceiling effects.

We treated each item as an ordinal indicator of the domain. The numerical value assigned to each category is irrelevant beyond its rank, for example, calling the lowest category 3 versus 18 makes no difference in how the item is treated or what the final score would be. This flexible approach does not make the strong assumption of a linear relationship between times and the underlying cognitive domain. The ordered categorical approach has much in common with spline approaches, which offer flexibility in modeling that may be important for constructs that may not be linear.

Missing data were a particular area of focus in our quality control effort. Some studies had little information about the reason for a missing data element. Other studies had specific codes, such as indicating participant refusal to complete an administered item or that the interviewer ran out of time so the item was not administered. After careful consideration, we decided to treat all types of missing data—regardless of codes available from the study—as if the item had not been administered.

Step A4: Confirmatory Factor Analyses

We modeled each domain separately using confirmatory factor analysis (CFA) with Mplus using robust weighted least squares, including terms for the mean and the variance (WLSMV) estimator (Beauducel & Herzberg, 2006; Flora & Curran, 2004).

As detailed in our prior article (Mukherjee et al. 2020), we consider several candidate bifactor structures. Our expert panel assigns subdomains based on theoretical considerations at the time we are considering domain assignments for each item. We also identify methods effects based on the ways items are administered. We perform agglomerative hierarchical cluster analyses to identify clusters of items with additional correlation structure. We then review proposed data-driven clusters of items with the expert panel and confirm that there is some thematic or methods based explanation for pairs or groups of items identified by the clustering algorithm; we only include secondary domains the expert panel agrees are plausible. When the more complicated model is consistent with theory (i.e., our content experts agree that the secondary domain structure makes theoretical sense), fits the data better (as evidenced by substantially better fit statistics), and produces substantially different scores (which we operationalize as differences greater than 0.3 logit units for at least 5% of the sample), we conclude that we need the more complicated model.

We used several criteria to compare these bifactor models for each domain, and found that in each case the agglomerative hierarchical clustering approach appeared to have the best fit, as detailed in Mukherjee et al. (2020).

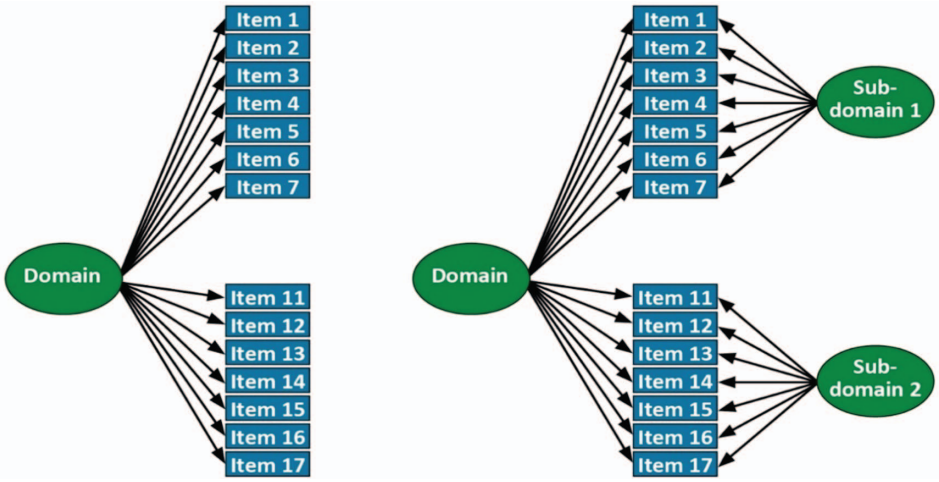
Once we had selected the best candidate bifactor model, we compared it with a single factor model, with no secondary structure (all items load only on the domain general factor).

We provide a schematic representation of single factor and bifactor models in Figure 2.

Our overall strategy in terms of single factor versus bifactor modeling was that we would choose the single factor model if

Figure 2

Single Factor (Left) and Bifactor (Right) Models of 14 Items From a Single Study



Note. The figure to the left depicts a single factor model of 14 items (1–7 and 11–17) that are depicted as loading on a single common factor. There are no secondary domains or residual covariances; this model forces all covariance between items to be captured by the single general factor (labeled "Domain" here). The figure to the right depicts the same 14 items, and a relationship with a general factor that captures covariance across all of the items. But different from the figure to the left, this bifactor model includes two subdomains (labeled "Subdomain 1" and "Subdomain 2"). These subdomains capture covariance among the subdomain items (e.g., Items 1–7 for Subdomain 1, and Items 11–17 for Subdomain 2) that is not shared with items outside that subdomain. A subdomain could be based on a methods effect (e.g., the same words from a word list learning task), or based on a common subset of a higher order domain (e.g., several items tapping set shifting in a model of executive functioning), or a data-driven subset based on agglomerative hierarchical clustering.

adding secondary factors did not markedly improve model fit and if adding secondary factors did not markedly impact any individual's score.

Our criteria for selecting the better model included evaluating fit statistics and concordance of model results with theory, such as positive loadings on secondary factors. The fit statistics we considered were the confirmatory fit index (CFI) where higher values indicate better fit; thresholds of 0.90 and 0.95 have been used in other settings as criteria for adequate or good fit (Hu & Bentler, 1999; Reeve et al., 2007); the Tucker–Lewis index, which has similar criteria as the CFI; and the root mean squared error of approximation (RMSEA), where lower values indicate better fit, and thresholds of 0.08 and 0.05 have been used in other settings as criteria for adequate or good fit (Hu & Bentler, 1999; Reeve et al., 2007).

When comparing the single factor model with the best bifactor model, we (a) determined whether loadings on the primary factor were within 10% of each other across the two models and (b) compared the scores for the single factor model versus scores for the bifactor model. We used as our threshold a difference of 0.30 units. We chose this value based on the default stopping rule for computerized adaptive testing; this has been used for years (S. W. Choi et al., 2010) as a default level of tolerable measurement imprecision. While arbitrary, this is a level of measurement imprecision that has been thought to be tolerable in a variety of situations. If there were a substantial number of people (typically 5%) for whom the differences in scores were larger than 0.3 from each other, and if the bifactor model conformed to our theory better and had better fit statistics, we selected the bifactor model as our choice for modeling a domain. Otherwise, we would select the simpler single factor model.

Step B1: Identification of Anchor Items

Cocalibration requires either the same people taking different tests or different tests sharing common items. Here we had common items. We identified candidate anchor items with identical content across tests administered in different studies and ensured that their relationship with the underlying ability tested was the same across studies by performing preliminary CFA models within each study, where we focused particularly on the pattern of loadings across the studies. Confirmed anchor items were then used to anchor the scales in each domain to a common metric. We show a depiction of candidate anchor items in Figure 3. We consulted a member of the expert panel (EHT) for anchor items selection review and confirmation.

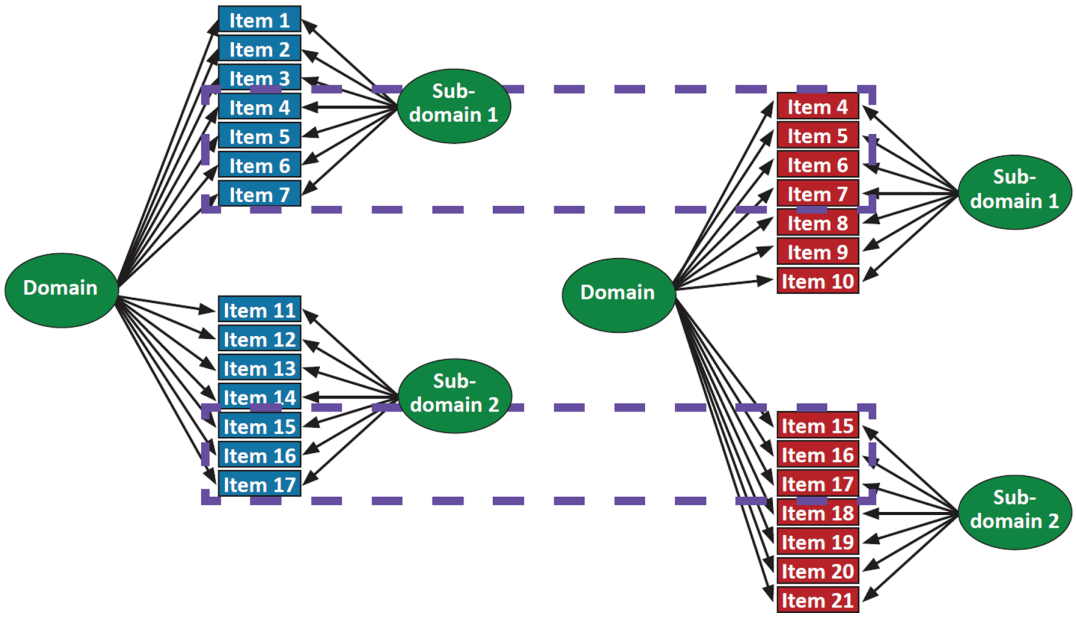
Step B2: Quality Control for Anchor Items

Anchor items were cleaned and recoded after considering item response data from all studies that administered the item, making sure that the range of responses to the anchor items was similar in each study. We carefully reviewed documentation from each study to ensure that the anchor item stimulus was precisely the same across studies, that the response options were precisely the same or could be recoded to be exactly equivalent across studies, and that we were mapping data from each study in a way that the same response would result in the same score regardless of the study in which the person was enrolled.

There were occasions where a potential candidate anchor item turned out to be administered in incompatible ways or scored in a way that could not be reconciled exactly across the two studies. These discrepancies were further discussed among the expert panel.

Figure 3

Data From Two Studies Illustrating Anchor Items



Note. This figure depicts data from a single domain for two studies. The blue study items are the same as those shown in Figure 2 in the bifactor model. The red study items appear to have some overlap, as depicted in the dashed blue boxes—red items 4–7 appear to be the same as blue items 4–7, and red items 15–17 appear to be the same as blue items 15–17. We pay close attention to these candidate anchor items, ensuring that the stimuli are identical and that the response coding is identical. The subset of items for which that turns out to be the case then are treated as anchor items, where the item parameters are forced to be the same between the blue study and the red study. Other items are treated as study-specific items, including those already understood to be study-unique (e.g., blue items 1–3 and 11–14, and red items 8–10 and 18–21).

If a candidate anchor item did not meet the approval of the panel, we included those items as indicators of the underlying domain in the different studies, but did not use those items as anchor items.

Step B3: Confirmatory Factor Analyses

We cocalibrated each cognitive domain by incorporating the components of the best model in each study (i.e., the final single-factor or bifactor model selected as described above) into one megacalibration model, as shown in Figure 4.

One particularly complicated aspect of cocalibrating scores using bifactor models is how to handle secondary domains. Some anchor items had loadings on the primary domain (e.g., memory) and on a secondary domain. That structure by itself does not lead to conceptual problems. Nevertheless, item representation of the secondary domain may vary across studies, with variable numbers of items, and potential missing data and identifiability issues. To address this, we used robust maximum likelihood (MLR) estimation that is robust to missing data, and if a secondary domain contained overlapping item(s) across studies along with study specific unique items, they were assigned to a common secondary domain in the megacalibration model. While the CFA model with the WLSMV estimator produces fit statistics in Mplus, the CFA model with the MLR estimator does not output fit statistics. We performed sensitivity analyses to confirm for ourselves that scores on the primary domain were minimally impacted by various ways of specifying the mean and variance on secondary domains. Since it made little difference how we specified these parameters, in our final models, we specified a mean of 0 and a variance of 1 for each secondary domain factor.

Once we had fit the final megacalibration model for each domain, we extracted item parameters (loadings and thresholds) for all items. These values then populated our item bank for each domain.

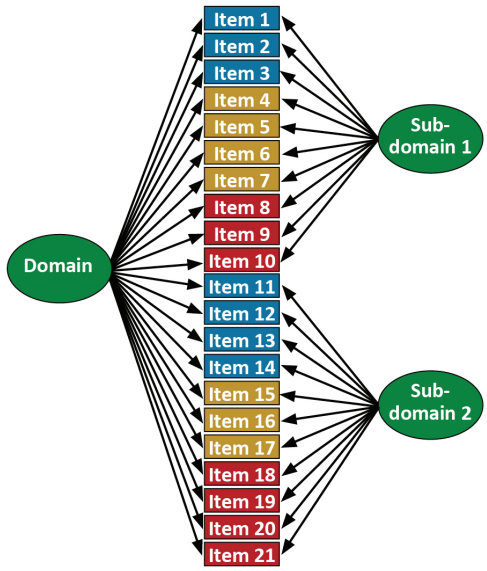
Scores From the Legacy Data Set for Each Time Point

We used each study's item parameters from the megacalibration model for a given domain (the item bank item parameters) to obtain scores for each person at each time point. We considered each study's data separately. We fixed all of the item parameters to their item bank values, and freely estimated means and variances for each factor in the model. Then we ran the model one additional time with all of the parameters fixed including the mean and the variance to extract factor scores for the primary factor (e.g., memory; labeled "Domain" in the figures) along with the corresponding standard errors. We used all participants with relevant data to obtain scores and standard errors for each domain, including people who may have been missing data entirely for some other domain.

Step C1: Domain Assignments for Subsequent Data Sets

Similar to the legacy data sets, our expert panel considers each element administered to study participants and categorize each one as an indicator of memory, executive functioning, language,

Figure 4
Cocalibration of the Red and Blue Studies



Note. Cocalibration model for data from Study 1 and Study 2. Study 1 data include blue and purple items, while Study 2 data include purple and red items. Beige items are anchors, which received extra attention and quality control (see above). This is referred to in this document as the "megacalibration model."

visuospatial, or none of these. Secondary domains are also assigned the same way as for the legacy data set described above.

Step C2: Data Quality Control on the Most Recent Study Visit Data

We consider data from the most recent study visit for each participant, as we did for the legacy studies. Some studies we have cocalibrated more recently have had cognitive batteries that have evolved over time and we have found it convenient to separate the data set into mutually exclusive subsets based on which cognitive batteries were administered. In essence, we treat each of these subsets as a distinct study that results in a less sparse covariance matrix of item responses and enables Mplus to estimate the item parameters and factor scores in a robust manner (see Scollard et al. article in this volume).

Step C3: Identify Candidate Anchor Items by Comparing Content With Domain Item Banks

We review each of the items from the new study and compare items with those already calibrated in the item bank. If content is identical and response options are identical, we consider the item to be a candidate anchor item. Procedures are the same as for the legacy data set described above.

Step C4: Quality Control for Anchor Items Added in Follow-Up Studies

We review distributions of responses in the study used to generate the item bank parameters and check that there is overlap with the distribution of item responses in the new study. We always recode data from the new study exactly as we did for the item bank study to ensure that the item is treated precisely the same way regardless of study. For some items the distribution of observed responses in the new study is sparse in some response categories, and this sparseness may persist after recoding. Since we are fixing parameters for anchor items as opposed to estimating parameters, modeling with sparseness in a response category will still work. What does not work is when there is a response category that is completely empty in the new study. To date this has consistently happened at the top or bottom category for an item. When this has happened we carefully excise the item parameters from that extreme and unobserved category so that the remaining parameters are appropriate for the observed distribution for that item. We take special care with this step as haste can lead to errors that could be difficult to catch; we pay particular attention to this step in our code review (see below).

Step C5: Confirmatory Factor Analysis Fixing Parameters for Anchor Item Banks to Item Bank Values

We used Mplus to analyze the most recent study visit data set. We fixed anchor items at their values from the item bank while new study specific items (nonanchors) were freely estimated. After this step, every item from the new study's cognitive battery has item parameters.

Step C6: Scores

We then fixed parameters for all of the items, and generated scores and corresponding standard errors for each person at each study visit. As for the legacy model this took two steps: first we freely estimated the mean and variance, and then we fixed the mean and variance to the estimated values to obtain individual scores and standard errors.

Step C7: Populate Item Bank With Any New Items From the New Study

The steps above can be used to generate scores as long as there is overlapping item content. Item parameters for nonanchors can be added to the item bank. We have had several data sets where we have determined that the distribution of ability levels observed in the new study was substantially different than that from our legacy studies. For example, we came across the Antiamyloid Treatment in Asymptomatic Alzheimer's (A4) study (Sperling et al., 2014) later on in our pipeline. Only cognitively intact people were considered for inclusion in the study, and only data from that screening visit were available for consideration. By design, there were no people with dementia in that data set, so lower portions of the ability distribution were not observed for any of our cognitive domains. We can obtain scores for such a data set assuming sufficient anchor item availability, but it would not be a good choice for calibrating item bank parameters for items first seen in that data set. A baseline data set from studies with constrained enrollment (i.e., studies like ACT) would not include people with poorer cognitive functioning. If a subsequent study included a broader range of participant ability levels because it included data from people with dementia, the items initially calibrated in the study of people without dementia would have truncated distributions which would lead to floor effects.

For studies where the full range of cognitive ability was observed, we update the item bank to incorporate item parameters from new study-specific items. In this way, the item banks for each domain continue to grow.

Code Review

Our quality control steps include a formal code review process (Vable et al., 2021) using GitHub (https://github.com). A primer on using Git and GitHub can be found at Blischak et al. (2016). In brief, we have created our private repository for the cocalibration effort and have folders designated for each study in our pipeline as well as relevant files related to our workflow. For a given study, as a team we choose a primary coder and a primary code reviewer for each cognitive domain a priori with a secondary code reviewer on standby if needed. We have three separate steps (precalibration step to look at factor structure; intermediate step to derive item parameters for unique items; derive scores) where a GitHub pull request is initialized by the primary coder and the review process involves reviews and updates until everyone approves it. GitHub makes it easier to track changes and one can go back or forward to any version of the code. The final code is pushed out making sure it is well annotated and reproducible with the primary coder and code reviewer's contact information for future use.

Studies Included in Legacy Model

The ACT Study

The ACT cohort is an urban and suburban elderly population randomly sampled from Kaiser Permanente-Washington (KPW) that includes 2,581 cognitively intact subjects age ≥65 who were enrolled between 1994 and 1998. An additional 811 subjects were enrolled in 2000–2002 using the same methods except oversampling clinics with more minorities. More recently, a continuous enrollment strategy was initiated in which new subjects are contacted, screened and enrolled to maintain a sample of 2000 people enrolled and at risk for dementia outcomes. This resulted in a total enrollment of 5,546 participants as of September 2018. Participants underwent assessment at study entry and every 2 years to evaluate cognitive function and collect demographic characteristics, medical history, health behaviors, and health status. In addition, information on participants' health care utilization and medication utilization were available from KPW electronic databases. Participants were assessed with the Cognitive Abilities Screening Instrument (CASI) at study entry and subsequent biennial visits (Teng et al., 1994). Participants with CASI scores ≤85 underwent a standardized diagnostic evaluation for dementia, including a physical and neurological examination, and additional neuropsychological tests (Kukull et al., 2002; Marcum et al., 2019). The extended neuropsychological battery includes tests such as WMS-R logical memory (2 stories), Mattis Dementia Rating Scale, Consortium to Establish a Registry for Alzheimer's Disease (CERAD) battery, Constructional Praxis, Verbal Paired Associates, Trails A & B, Clock Drawing, Boston Naming Test (BNT), and verbal fluency measures. All of these data are reviewed at consensus conference where research criteria for dementia and Alzheimer's disease are determined.

The ADNI Study

ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations. Study resources and data are available through its website (http://adni.loni.usc.edu). The initial 5-year study (ADNI1) was extended by 2 years in 2009 (ADNIGO), and in 2011 and 2016 by further competitive renewals (ADNI2 and ADNI3). Through April of 2020, 3,016 individuals were enrolled across the different ADNI waves. The study was conducted after institutional review board approval at each site. Written informed consent was obtained from study participants or authorized representatives. Additional details of the study design are available elsewhere (Weiner et al., 2010, 2017). ADNI's neuropsychological battery included the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Schedule–Cognition (ADAS-Cog), BNT, Rey Auditory Verbal Learning Test, Wechsler Memory Scale–Revised (WMS-R) Digit Span, WMS-R Logical Memory, Trails A & B, clock drawing, and animal- and (for ADNI1 only) vegetable fluency. ADNI administered Montreal Cognitive Assessment (MoCA) items beginning in ADNIGO.

The ROS Study

The ROS has been ongoing since 1993, with a rolling admission. Through February of 2020, 1,456 older nuns, priests, and brothers from across the United States initially free of dementia who agreed to annual clinical evaluation and brain donation at the time of death completed their baseline evaluation.(Bennett, Schneider, Arvanitakis, et al., 2012)

The MAP Study

The MAP has been ongoing since 1997, also with a rolling admission. Through February of 2020, 2,163 older persons from across northeastern Illinois initially free of dementia who agreed to annual clinical evaluation and organ donation at the time of death completed their baseline evaluation.(Bennett, Schneider, Buchman, et al., 2012; Bennett et al., 2005)

ROS/MAP administers 21 cognitive tests such as CERAD test, MMSE, East Boston Story, logical memory story from WMS-R, BNT, semantic fluency measures, WMS-R Digit Span, Symbol Digit Modalities Test, Judgment of Line Orientation, Standard Progressive Matrices, and Number Comparison (Bennett et al., 2018; Wilson et al., 2002). This comprehensive battery overlaps mostly across ROS and MAP (19 out of 21) and enables investigation of episodic memory, semantic memory, working memory, perceptual speed, and visuospatial functioning.

Transparency and Openness

We report the variables used in each study, how we determined our sample size, all data exclusions, all analyses, and all measures in the study. All analysis scripts are available from authors on request and all cognitive data and the harmonized cognitive domains used can be requested from the parent studies. Data were analyzed using Mplus v7.4 and Stata v16. The analyses conducted in this article were not preregistered.

Results

Findings From the Legacy Data Sets

We included $n = 5{,}546$ from ACT, $n = 3{,}016$ from ADNI, $n = 1{,}456$ from ROS, and $n = 2{,}163$ from MAP in our legacy cocalibration model. Demographic and clinical characteristics from the most recent study visit are shown in Table 1. We fixed the mean at 0 and variances at 1 for the primary and secondary domains to estimate item parameters. We freely estimated the mean and variances of the primary and secondary factors when running domain-specific models to derive scores in each study.

Legacy Study Items in the Item Bank for Each Domain

For the memory domain, MMSE orientation items and logical memory immediate and delayed recall were administered in each of the studies and served as anchor items. The ROS and MAP battery added an additional 13 items to the item bank, the ACT study added 25, and ADNI added 20 more (Supplemental Tables 1–4). For executive functioning, ACT and ADNI had Trails A and B in common, and ADNI, ROS, and MAP had digit span forward and backward and the WORLD backwards item from the MMSE. ROS and MAP added four additional items, ACT added eight items, and ADNI added seven items from all waves plus seven from the MoCA in later waves of the ADNI study (Supplemental Tables 5–8). For language, all four studies had the reading and command items from the MMSE as well as animal fluency in common, ACT, ROS, and MAP had the 15-item version of the Boston Naming Test in common, and ADNI, ROS, and MAP had repeating a phrase and writing a sentence from the MMSE in common. ROS and MAP added 11 additional items, ACT added eight additional items, and ADNI added five additional items plus six from the MoCA in later waves of the ADNI study (Supplemental Tables 9–12). For visuospatial functioning, all four studies had interlocking pentagons, ROS and MAP added the Judgment of Line Orientation, ACT added five additional items and ADNI added six additional items (Supplemental Tables 13–16).

With the most recent data pulls, we derived scores for 5,546 individuals from ACT where each individual had all four scores for 98% of their visits ($n = 26{,}498$ scores). In ADNI, we have scores for 3,189 individuals where we have all four scores for 90% of their visits ($n = 11{,}680$). In ROS, we derived scores for 1,456 individuals where all four scores were present for 94% of the observations ($n = 14{,}805$). In MAP, we derived scores for 2,163 individuals where all four scores were present for 96% of the observations ($n = 14{,}350$). Distributions of these scores in each of the four studies are shown in Figure 5.

Table 1
Demographic and Clinical Characteristics of the Legacy Studies at the Most Recent Study Visit

Variable	ACT	ADNI	ROS	MAP
Sample size, *n*	5,546	3,016	1,456	2,163
Age, mean (*SD*)	81.9 (7.8)	74.9 (8.7)	85.8 (7.4)	86.0 (7.9)
Female, (%)	58.2	48.1	71.6	73.5
Education, mean (*SD*)	14.9 (3.2)	16.1 (2.8)	18.4 (3.3)	14.9 (3.3)
Self-reported race, %				
Non-Latinx White	88.8	88.1	89.6	88.5
African/American	3.6	4.6	5.7	5.2
Hispanic/Latino	1.1	3.7	4.3	5.5
Others	6.5	3.6	0.4	0.8
Cognitive diagnosis at most recent visit, %				
Cognitively normal	92.2	38.3	42.2	50.4
Mild Cognitive Impairment (MCI)	N/A[a]	32.0	23.0	23.7
Diagnosed with AD	6.0	29.7	33.4	24.5
Other dementia	1.8	N/A	1.4	1.4

Note. ACT = adult changes in thought; ADNI = Alzheimer's Disease Neuroimaging Initiative; ROS = Religious Orders Study; MAP = Memory and Aging Project; AD = Alzheimer's disease; *SD* = standard deviation. A few individuals were missing some of these demographic characteristics.
[a] In ACT, MCI as a diagnosis is generally not made.

Follow-Up Study 1: Findings From the Rush Minority Aging Research Study Data

The Minority Aging Research Study (MARS) is a longitudinal, epidemiologic cohort study of decline in cognitive function and risk of Alzheimer's disease (AD) in older African Americans, with brain donation after death added as an optional component for those willing to consider organ donation (Barnes et al., 2012). A comprehensive neuropsychological battery of 23 cognitive tests is administered at each annual visit. The tests we used for cocalibration overlapped completely with what we had seen in ROS and MAP, which were part of the legacy model. As a result, all items were anchors and we were able to directly use all our derived item parameters to obtain scores for all MARS participants across all time points.

We derived scores for 767 individuals from MARS where each individual had all four scores for 97% of their visits ($n = 5{,}075$ scores). Demographic and clinical characteristics at most recent visit are shown in Table 2. Violin plots for each domain are shown in Figure 6.

Follow-Up Study 2: Findings From the National Alzheimer's Coordinating Center Data

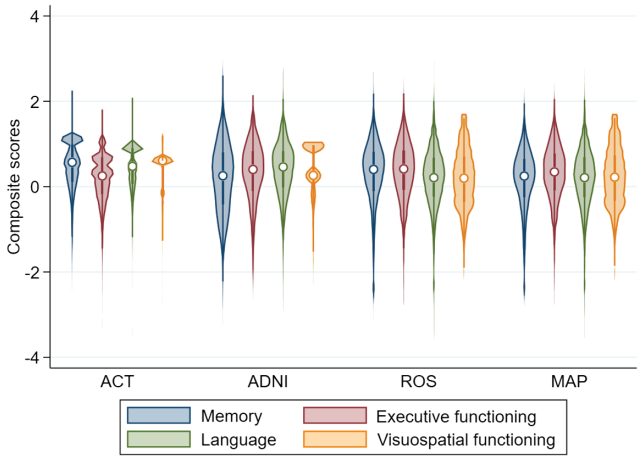
The National Alzheimer's Coordinating Center (NACC) is responsible for developing and maintaining a database of participant information collected from Alzheimer's Disease Centers (ADCs) funded by the National Institute on Aging (NIA; Beekly et al., 2007). The neuropsychological test battery from the Uniform Data Set (UDS) of the Alzheimer's Disease Centers (ADC) program of the National Institute on Aging consists of brief measures of attention, processing speed, executive functioning, episodic memory, and language (Weintraub et al., 2018, 2009). The UDS battery has evolved over time from Version 1.0 to 2.0 to 3.0.

We included individuals with baseline age ≥ 60 for cocalibration purpose. Demographic and clinical characteristics from the most recent study visit are shown in Table 3.

By design there was substantial overlap in the battery used for the ADNI study and for the NACC data. We considered UDS data in two batches, essentially as two separate studies. The only visuospatial item administered in the UDS is the dichotomous interlocking pentagons item, which was not sufficient to obtain scores for that domain. We were able to obtain cocalibrated scores for the other domains. Beyond the pentagons item, NACC collects other MMSE items as composites. For example, the five orientation to time items from the MMSE are reported to NACC as a single score. For these situations, we reran the legacy model with everything other than these composites treated as anchor items, obtaining item parameters on the same metric for the MMSE composite scores from the legacy

Figure 5

Violin Plot of the Distributions for Each of the Cognitive Scores Across All Time Points by Study Used in Legacy Model



Note. The violin plot displays the median as a circle, the first-to-third interquartile range as a narrow, shaded box, and the lower-to-upper adjacent value range as a vertical line. The violins are mirrored density curves. ACT = adult changes in thought; ADNI = Alzheimer's Disease Neuroimaging Initiative; ROS = Religious Orders Study; MAP = Memory and Aging Project.

data. We then used these item parameters for the MMSE composite scores in the NACC data, along with other anchors as shown in Supplemental Tables 17–19. More details about this process can be found in Supplemental Text 2. We derived scores for 41,459 individuals from NACC UDS 1/2/3 where each individual had all four scores for 87% of their visits (total = 145,028). The distribution of scores is shown in Figure 7.

Other Data Set Cocalibrated and Harmonized

With a growing item bank, we have been able to cocalibrate and harmonize cognitive domains from various aging studies such as the A4 Study, the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging (Ellis et al., 2009), the Baltimore Longitudinal Study of Aging (BLSA; Ferrucci, 2008), and the Framingham Heart Study (FHS; Elias et al., 1995). The cognitive scores can be obtained from the parent studies via data user agreement (DUA). Taken together, across all of these studies, we have cocalibrated cognitive data for 76,723 individuals from 10 studies.

Table 2

Demographic and Clinical Characteristics of MARS at the Most Recent Study Visit

Variable	MARS
Sample size, *n*	767
Age, mean (*SD*)	79.9 (7.3)
Female, (%)	77.2
Education, mean (*SD*)	14.8 (3.5)
Self-reported race, %	
Non-Latinx White	0.0
African/American	99.9
Hispanic/Latino	0.0
Others	0.1
Cognitive diagnosis at most recent visit, %	
Cognitively normal	74.8
MCI	21.0
Diagnosed with AD	3.9
Other dementia	0.3

Note. MARS = Minority Aging Research Study; *SD* = standard deviation; MCI = Mild Cognitive Impairment. Percentages for cognitive diagnosis (DX) shown for nonmissing data only.

Discussion

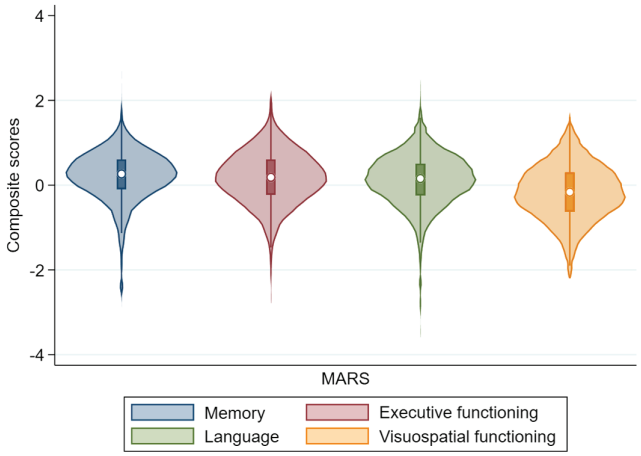
We cocalibrated cognitive data across multiple studies of older adults using a modern psychometrics approach. This approach, which is well-suited to our purpose, was easily adapted from its application to educational settings, to the specific challenges from cognitive testing of older adults.

Our expert panel categorized each item as best reflecting single cognitive domain, but for several items also identified a second domain that the item also tapped. We used CFA to assess whether the items that best reflected a domain load well for that domain. We wanted each domain to contain a mutually exclusive set of items, and as a result, did not explore factor analysis models for domains that included items assigned to their secondary domains. For genetic analyses, one of the motivating use cases for our harmonization efforts, there is tremendous interest in pleiotropy, where a particular genetic factor may underlie multiple phenotypes. Allowing cross-loading of a single item on multiple domains would induce correlation between domain scores and would make evaluation of pleiotropy findings at least difficult if not impossible (Solovieff et al., 2013). Others with different goals could have made different modeling choices.

One contribution we make in this article is that we used bifactor models to cocalibrate these data. As shown, the introduction of secondary factors requires careful thought and consideration. We compared several methods of deriving candidate secondary structures. We compared several methods of deriving candidate secondary structures. While the different bifactor models produced consistent results—suggesting some robustness to the specification of the secondary factors—it should also be emphasized that bifactor models had substantially better fit than single factor models, and that the bifactor models and single factor models produced scores that were substantially different from each other for some people. In many cases in our workflow we are faced with the overarching question of whether we need a more complicated bifactor model or whether a simpler single factor model would be "good enough." On the other hand, even if a more complicated model is consistent with theory, if fit statistics are either marginally better, very similar, or worse, and if the scores from a more complicated model do not substantially differ from those of a simpler model, we would choose the simpler model. In this instance, application of that approach led us to choose bifactor models rather than single factor models. But in the case of the sensitivity analyses of different choices we could make for subdomains, we did not find evidence that we needed a different model. Both of these sets of results can be seen as examples of the same overarching strategy. In each case, fit statistics led us to choose models where the secondary structure was derived from agglomerative hierarchical clustering. The resulting scores for each domain account for these secondary data structures, which essentially avoids overemphasizing responses that would otherwise by somewhat too influential on the overall score.

Figure 6

Violin Plot Showing Distribution of Scores Across All Time Points for Four Cognitive Domains in MARS



Note. MARS = Minority Aging Research Study.

We used data from the most recent study visit for each person to calibrate items. This strategic choice ensured that each individual would only be included once in our calibration modeling, so we did not have to address within-person correlations. This choice also maximized the spread of observed ability levels for each domain, which is desirable for a calibration sample. Some data sets by design were characterized by constrained variation in one or more domains. For example, A4 included the baseline data point from a group of cognitively normal older adults, which meant there were no people with dementia and no substantially impaired scores. For these data sets, we can obtain scores that are cocalibrated based on the inclusion of anchor items, but we did not use study-specific item parameters from these data sets in our item bank. All of the items in our item bank had parameters estimated from samples with a broad range of ability levels.

Table 3

Demographic and Clinical Characteristics for Individuals With Study Baseline Age ≥60 in the NACC Data Set at the Most Recent Study Visit

Variable	UDS 1 & 2	UDS 3
Sample size, *n*	29,154	15,232
Age, mean (*SD*)	77.1 (8.4)	76.3 (8.3)
Female, (%)	56.8	58.4
Education, mean (*SD*)	14.9 (3.6)	15.8 (3.2)
Self-reported race, %		
Non-Latinx White	80.5	80.1
African/American	14.3	14.1
Hispanic/Latino	1.1	0.7
Others	4.1	5.1
Cognitive diagnosis[a] at most recent visit, %		
Cognitively Normal	33.8	48.0
MCI	19.2	15.7
Diagnosed with AD	44.0	34.3
Other dementia	3.0	2.0

Note. NACC = National Alzheimer's Coordinating Center; MCI = Mild Cognitive Impairment; AD = Alzheimer's disease; UDS = uniform data set; *SD* = standard deviation.
[a] Based on primary, contributing, or noncontributing cause Alzheimer's disease.

Of note, we used similar methods in previous work (Mukherjee et al., 2020). There are important differences here. First and foremost, that work focused exclusively on samples of people with clinical Alzheimer's dementia, and all recoding and model calibration was performed on those data. In the present work we include people across the entire range of the cognitive ability scale, from completely unimpaired to severely impaired. Our work for this article is thus applicable to people at all levels of cognitive ability, not limited to people with clinical Alzheimer's dementia.

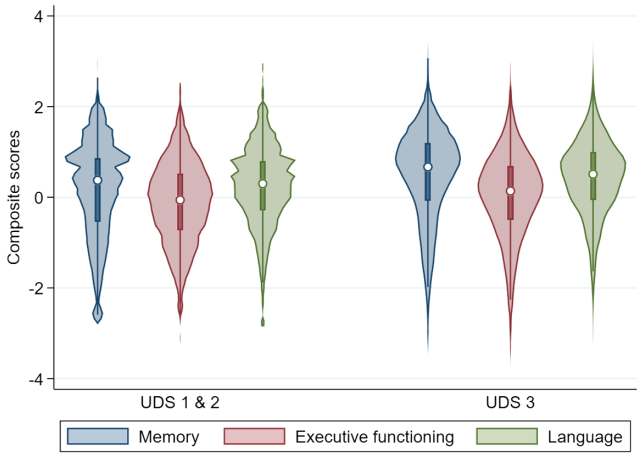
As in any item banking effort, anchor items are essential for successfully linking scores across different studies. We pay close attention to anchor items as detailed here, ensuring that the stimuli are identical, that the responses are scored in an identical fashion, and that the distribution of observed scores has substantial overlap across studies.

The cocalibration approaches described in this article will enable investigations of associations with late life cognitive functioning and decline using data from multiple studies, even though those studies measured cognition in older adults using different neuropsychological tests. The payoff for the work we have done is the ease of use of the resulting scores. They address important psychometric challenges in the parent data, so the user of the scores can focus on their scientific questions of interest.

This article has focused on considerations in cocalibrating scores across studies that used different batteries. We did not address validity. There are many layers that ensure the validity of our cocalibrated scores. First, these scores are derived from cognitive tests administered by prominent studies that have had their methods peer reviewed many times. The modern psychometrics approach we used does not diminish the validity of the underlying measures. Second, our approach to domain assignments began with our expert panel, who in turn are guided by disciplinary considerations in the field of neuropsychology. The tests whose items we analyzed here have been widely used, producing a vast literature in applied settings. Furthermore, we are transparent with our choices and indeed present our domain assignments to the scientific community in this article. Others could assign items to different domains. We suspect that differences in assignment across content experts likely would reflect matters of degree. For example, we assigned the overlapping pentagons item to the visuospatial domain, though certainly there are aspects of executive functioning that are required to successfully complete this item, and it could be argued that item would be a better representative of the executive functioning domain than the visuospatial domain. Even in such an instance, however, we suspect such a content expert would agree with our panel that the interlocking pentagons item is also an indicator of the visuospatial domain. Disagreements of this sort on matters of degree do not rise to the level of challenging the overall validity of any domain score. The only real such challenge to overall validity would be if an item simply was not an indicator of the domain our experts assigned it to, which we think has not happened. Third, we have in previous work compared modern psychometric scores alongside classical test theory-derived scores for the same domain, using a variety of validity comparisons including known group comparisons, strength of association with a priori selected imaging findings, ability to predict decline over time and conversion from Mild Cognitive Impairment (MCI) to AD (S. E. Choi et al., 2020; Crane et al., 2012; L. E. Gibbons et al., 2012).

Figure 7

Violin Plot Showing Distribution of Scores Across All Time Points for Three Cognitive Domains in UDS 1 & 2 and UDS 3



Note. UDS = uniform data set.

The cocalibration we do here is a minor tweak of the calibration we have done previously and evaluated validity with, using essentially the same modeling strategy. There is a simple practical reason we do not provide additional novel analyses of the validity of the cocalibrated scores, which is that given the challenges we had to address to develop cocalibrated scores, there is no classical test theory-derived approach for harmonizing these data that we would recommend. All such methods we are aware of make assumptions that are not supported by data. For example, if we took *z*-scores within a domain for each study, we would not have a way to link studies together without making a huge assumption that the means and standard deviations of the two samples are exactly the same. As shown elsewhere in this issue (Hampton et al., 2020), when we have evaluated this assumption across studies we have not found it to be plausible. Standard approaches widely used in the field such as *z*-scores make strong assumptions that must be correct for resulting scores to be valid (McNeish & Wolf, 2020). Some studies change all or part of the neuropsychological battery over time and it becomes impossible to cocalibrate cognitive data with naïve total score and *z*-score approaches and derive scores on the same scale. There are other approaches such as linear linking for related traits (Nichols et al., 2021) built using item response theory machinery that can be used for cocalibration but it uses additional assumptions and is not amenable to domains with secondary data structure. Our approaches make far fewer assumptions; at each step, as outlined here, we have made careful modeling choices that are consistent with the data. A limitation of this field is that there is no robust metric to get a sense of cocalibration accuracy.

One limitation of our current workflow is that our choice of which datasets to begin our procedures with was based in part on convenience, specifically which data sets we had access to at the beginning of our work. As more data sets become available, it will become possible to consider the implications of making different choices. Investigations of those choices may be very useful in determining the potential impact of initial selection of studies or pooling all of the studies together, as well as the cumulative impact of sequential item parameter instabilities.

We do not incorporate methods to account for repeated measures in our CFAmodels. In this initial work, we chose a single measurement occasion for each individual. There could be some learning effects that could have an impact on item difficulty or discrimination. Further work will be needed to investigate this issue. We are somewhat comforted that, in many instances, intervals between testing are many months (ADNI and others), a year (ROS/MAP and others), or even 2 years (ACT and others) apart; retest or learning effects are thought to be more salient with study visits that are close to each other. This is an active area of research (Jutten et al., 2020). Another limitation is that we didn't perform any formal Differential Item Functioning (DIF) testing across the suite of studies. DIF occurs when groups (such as defined by sex, ethnicity, age, or education) have different probabilities of endorsing a given item after controlling for overall scores. We plan to examine and adjust our scores for DIF in future studies, especially across ethnicity (e.g., MAP and MARS), and adjust for it if it turns out to be impactful (Crane et al., 2007; Dmitrieva et al., 2015).

To date these efforts have enabled us to cocalibrate hundreds of thousands of scores from tens of thousands of individual study participants. These rich data are available for interested investigators to use. The item parameters we have generated to date are stored and deriving scores for additional study participants and observations becomes a much simpler task in subsequent waves from ongoing studies and indeed for new studies with overlapping content. There are multiple ongoing funded initiatives that have or will use these cocalibrated cognitive data. Cocalibrated cognitive scores were used to derive a measure of resilience (Dumitrescu et al., 2020) facilitating meta-analysis of genetic results across cohorts enabling us to find candidate loci associated with resilience. These analyses would not have been possible without the cocalibrated scores; cocalibrated scores facilitated analyses of the replicability of genotype–phenotype association signals across multiple studies that used different instruments to measure cognition. The Alzheimer's disease genetics community has seen value in the approaches we have taken; we propose to use these same approaches in the now funded U24 AG074855, "Alzheimer's Disease Sequencing Project Phenotype Harmonization Consortium." Our hope is that this protocol article will serve a valuable role in all these initiatives in documenting our workflow for cocalibrating cognitive scores across studies.

References

Barnes, L. L., Shah, R. C., Aggarwal, N. T., Bennett, D. A., & Schneider, J. A. (2012). The Minority Aging Research Study: Ongoing efforts to obtain brain donation in African Americans without dementia. *Current Alzheimer Research*, *9*(6), 734–745. https://doi.org/10.2174/156720512801322627

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., Hubbard, J. L., Koepsell, T. D., Morris, J. C., Kukull, W. A., & the NIA Alzheimer's Disease Centers. (2007). The National Alzheimer's Coordinating Center (NACC) database: The Uniform Data Set. *Alzheimer Disease and Associated Disorders*, *21*(3), 249–258. https://doi.org/10.1097/WAD.0b013e318142774e

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., & Schneider, J. A. (2018). Religious orders study and Rush Memory and

Aging Project. *Journal of Alzheimer's Disease*, *64*(S1), S161–S189. https://doi.org/10.3233/JAD-179939

Bennett, D. A., Schneider, J. A., Arvanitakis, Z., & Wilson, R. S. (2012). Overview and findings from the Religious Orders Study. *Current Alzheimer Research*, *9*(6), 628–645. https://doi.org/10.2174/15672051280132 2573

Bennett, D. A., Schneider, J. A., Buchman, A. S., Barnes, L. L., Boyle, P. A., & Wilson, R. S. (2012). Overview and findings from the Rush Memory and Aging Project. *Current Alzheimer Research*, *9*(6), 646–663. https:// doi.org/10.2174/156720512801322663

Bennett, D. A., Schneider, J. A., Buchman, A. S., Mendes de Leon, C., Bienias, J. L., & Wilson, R. S. (2005). The Rush Memory and Aging Project: Study design and baseline characteristics of the study cohort. *Neuroepidemiology*, *25*(4), 163–175. https://doi.org/10.1159/000087446

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A quick introduction to version control with git and gitHub. *PLoS Computational Biology*, *12*(1), Article e1004668. https://doi.org/10.1371/journal.pcbi.1004668

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10 .1017/CBO9780511490026

Choi, S. E., Mukherjee, S., Gibbons, L. E., Sanders, R. E., Jones, R. N., Tommet, D., Mez, J., Trittschuh, E. H., Saykin, A., Lamar, M., Rabin, L., Foldi, N. S., Sikkes, S., Jutten, R. J., Grandoit, E., Mac Donald, C., Risacher, S., Groot, C., Ossenkoppele, R., … Crane, P. K. (2020). Development and validation of language and visuospatial composite scores in ADNI. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *6*(1), Article e12072. https://doi.org/10.1002/trc2.12072

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *70*(6), 1–17. https://doi.org/10.1177/0013164410387338

Crane, P. K., Carle, A., Gibbons, L. E., Insel, P., Mackin, R. S., Gross, A., Jones, R. N., Mukherjee, S., Curtis, S. M., Harvey, D., Weiner, M., Mungas, D., & the Alzheimer's Disease Neuroimaging Initiative. (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, *6*(4), 502–516. https://doi.org/10.1007/s11682-012-9186-z

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R. D., & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *16*(Suppl. 1), 69–84. https://doi.org/10.1007/s11136-007-9185-5

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, *61*(10), 1018–1027. https://doi.org/10.1016/j.jclinepi.2007.11.011

Dmitrieva, N. O., Fyffe, D., Mukherjee, S., Fieo, R., Zahodne, L. B., Hamilton, J., Potter, G. G., Manly, J. J., Romero, H. R., Mungas, D., & Gibbons, L. E. (2015). Demographic characteristics do not decrease the utility of depressive symptoms assessments: Examining the practical impact of item bias in four heterogeneous samples of older adults. *International Journal of Geriatric Psychiatry*, *30*(1), 88–96. https:// doi.org/10.1002/gps.4121

Dumitrescu, L., Mahoney, E. R., Mukherjee, S., Lee, M. L., Bush, W. S., Engelman, C. D., Lu, Q., Fardo, D. W., Trittschuh, E. H., Mez, J., Kaczorowski, C., Hernandez Saucedo, H., Widaman, K. F., Buckley, R., Properzi, M., Mormino, E., Yang, H. S., Harrison, T., Hedden, T., … Hohman, T. J. (2020). Genetic variants and functional pathways associated with resilience to Alzheimer's disease. *Brain: A Journal of Neurology*, *143*(8), 2561–2575. https://doi.org/10.1093/brain/awaa209

Elias, M. F., D'Agostino, R. B., Elias, P. K., & Wolf, P. A. (1995). Neuropsychological test performance, cognitive functioning, blood pressure, and age: The Framingham Heart Study. *Experimental Aging Research*, *21*(4), 369–391. https://doi.org/10.1080/03610739508253991

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., … the AIBL Research Group. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. https://doi.org/10.1017/S1041610209009405

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Ferrucci, L. (2008). The Baltimore Longitudinal Study of Aging (BLSA): A 50-year-long journey and plans for the future. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *63*(12), 1416–1419. https://doi.org/10.1093/gerona/63.12.1416

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X .9.4.466

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https:// doi.org/10.1016/0022-3956(75)90026-6

Gatz, M., Reynolds, C. A., Finkel, D., Hahn, C. J., Zhou, Y., & Zavala, C. (2015). Data harmonization in aging research: Not so fast. *Experimental Aging Research*, *41*(5), 475–495. https://doi.org/10.1080/0361073X.2015 .1085748

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S. M., Mungas, D., Crane, P. K., & the Alzheimer's Disease Neuroimaging Initiative. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, *6*(4), 517–527. https://doi.org/10.1007/s11682-012-9176-1

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19. https://doi.org/10 .1177/0146621606289485

Gross, A. L., Tommet, D., D'Aquila, M., Schmitt, E., Marcantonio, E. R., Helfand, B., Inouye, S. K., Jones, R. N., & the BASIL Study Group. (2018). Harmonization of delirium severity instruments: A comparison of the DRS-R-98, MDAS, and CAM-S using item response theory. *BMC Medical Research Methodology*, *18*(1), Article 92. https://doi.org/10 .1186/s12874-018-0552-4

Gruhl, J., Erosheva, E. A., & Crane, P. K. (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *The Annals of Applied Statistics*, *7*(4), 2361–2383. https://doi.org/10.1214/13-AOAS675

Hall, K. S., Gao, S., Emsley, C. L., Ogunniyi, A. O., Morgan, O., & Hendrie, H. C. (2000). Community screening interview for dementia (CSI 'D'); performance in five disparate study sites. *International Journal of Geriatric Psychiatry*, *15*(6), 521–531. https://doi.org/10.1002/1099-1166 (200006)15:6<521::AID-GPS182>3.0.CO;2-F

Hall, K. S., Hendrie, H. C., Brittain, H. M., & Norton, J. A. (1993). Development of a dementia screening interview in two distinct languages. *International Journal of Methods in Psychiatric Research*, *2*, 1–28.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hampton, O. L., Mukherjee, S., Properzi, M. J., Schultz, A. P., Crane, P. K., Gibbons, L. E., Hohman, T. J., Maruff, P., Lim, Y. Y., Amariglio, R. E., Papp, K. V., Johnson, K. A., Rentz, D., Sperling, R. A., & Buckley, R. F. (2020). Harmonizing the preclinical Alzheimer cognitive composite for

multi-cohort studies. *Alzheimer's & Dementia*, *16*(Suppl. 9), Article e047423. https://doi.org/10.1002/alz.047423

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S. E., Lamar, M. L., Louden, D. K. N., Rich, J., Tommet, D., Crane, P. K., & Rabin, L. A. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: A literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *12*(1), Article e12055. https://doi.org/10.1002/dad2.12055

Kukull, W. A., Higdon, R., Bowen, J. D., McCormick, W. C., Teri, L., Schellenberg, G. D., van Belle, G., Jolley, L., & Larson, E. B. (2002). Dementia and Alzheimer disease incidence: A prospective cohort study. *Archives of Neurology*, *59*(11), 1737–1746. https://doi.org/10.1001/archneur.59.11.1737

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. https://doi.org/10.1177/0146621605275414

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores, with contributions by Allan Birnbaum*. Addison-Wesley.

Marcum, Z. A., Walker, R. L., Jones, B. L., Ramaprasan, A., Gray, S. L., Dublin, S., Crane, P. K., & Larson, E. B. (2019). Patterns of antihypertensive and statin adherence prior to dementia: Findings from the adult changes in thought study. *BMC Geriatrics*, *19*(1), Article 41. https://doi.org/10.1186/s12877-019-1058-6

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Montine, T. J., Sonnen, J. A., Montine, K. S., Crane, P. K., & Larson, E. B. (2012). Adult changes in thought study: Dementia is an individually varying convergent syndrome with prevalent clinically silent diseases that may be modified by some commonly used therapeutics. *Current Alzheimer Research*, *9*(6), 718–723. https://doi.org/10.2174/156720512801322555

Mukherjee, S., Mez, J., Trittschuh, E. H., Saykin, A. J., Gibbons, L. E., Fardo, D. W., Wessels, M., Bauman, J., Moore, M., Choi, S. E., Gross, A. L., Rich, J., Louden, D. K. N., Sanders, R. E., Grabowski, T. J., Bird, T. D., McCurry, S. M., Snitz, B. E., Kamboh, M. I., … Crane, P. K. (2020). Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Molecular Psychiatry*, *25*(11), 2942–2951. https://doi.org/10.1038/s41380-018-0298-8

Muthen, L. K., & Muthen, B. O. (1998–2012). *Mplus user's guide* (7th ed.).

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nichols, E. L., Cadar, D., Lee, J., Jones, R. N., & Gross, A. L. (2021). Linear linking for related traits (LLRT): A novel method for the harmonization of cognitive domains with no or few common items. *Methods*. Advance online publication. https://doi.org/10.1016/j.ymeth.2021.11.011

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D., & the PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Suppl. 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495. https://doi.org/10.1038/nrg3461

Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: Stopping AD before symptoms begin? *Science Translational Medicine*, *6*(228), Article 228fs13. https://doi.org/10.1126/scitranslmed.3007941

Teng, E. L., & Chui, H. C. (1987). The Modified Mini-Mental State (3MS) examination. *The Journal of Clinical Psychiatry*, *48*(8), 314–318.

Teng, E. L., Hasegawa, K., Homma, A., Imai, Y., Larson, E., Graves, A., Sugimoto, K., Yamaguchi, T., Sasaki, H., Chiu, D., & White, L. R. (1994). The Cognitive Abilities Screening Instrument (CASI): A practical test for cross-cultural epidemiological studies of dementia. *International Psychogeriatrics*, *6*(1), 45–58. https://doi.org/10.1017/S1041610294001602

Vable, A. M., Diehl, S. F., & Glymour, M. M. (2021). Code review as a simple trick to enhance reproducibility, accelerate learning, and improve the quality of your team's research. *American Journal of Epidemiology*, *190*(10), 2172–2177. https://doi.org/10.1093/aje/kwab092

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Weiner, M. W., Aisen, P. S., Jack, C. R., Jr., Jagust, W. J., Trojanowski, J. Q., Shaw, L., Saykin, A. J., Morris, J. C., Cairns, N., Beckett, L. A., Toga, A., Green, R., Walter, S., Soares, H., Snyder, P., Siemers, E., Potter, W., Cole, P. E., & Schmidt, M. (2010). The Alzheimer's disease neuroimaging initiative: Progress report and future plans. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *6*(3), 202–211.e207. https://doi.org/10.1016/j.jalz.2010.03.007

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jr., Jagust, W., Morris, J. C., Petersen, R. C., Salazar, J., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & the Alzheimer's Disease Neuroimaging Initiative. (2017). The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia*, *13*(5), 561–571. https://doi.org/10.1016/j.jalz.2016.10.006

Weintraub, S., Besser, L., Dodge, H. H., Teylan, M., Ferris, S., Goldstein, F. C., Giordani, B., Kramer, J., Loewenstein, D., Marson, D., Mungas, D., Salmon, D., Welsh-Bohmer, K., Zhou, X. H., Shirk, S. D., Atri, A., Kukull, W. A., Phelps, C., & Morris, J. C. (2018). Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Disease and Associated Disorders*, *32*(1), 10–17. https://doi.org/10.1097/WAD.0000000000000223

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., Cummings, J., DeCarli, C., Foster, N. L., Galasko, D., Peskind, E., Dietrich, W., Beekly, D. L., Kukull, W. A., & Morris, J. C. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): The neuropsychologic test battery. *Alzheimer Disease and Associated Disorders*, *23*(2), 91–101. https://doi.org/10.1097/WAD.0b013e318191c7dd

Wilson, R. S., Beckett, L. A., Barnes, L. L., Schneider, J. A., Bach, J., Evans, D. A., & Bennett, D. A. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychology and Aging*, *17*(2), 179–193. https://doi.org/10.1037/0882-7974.17.2.179

Received October 30, 2021
Revision received April 22, 2022
Accepted May 2, 2022 ■