ADSP Phenotype Harmonization Consortium – Derivation of Cognitive Composite Scores

Shubhabrata Mukherjee[1], Seo-Eun Choi[1], Michael L. Lee[1], Phoebe Scollard[1], Emily H. Trittschuh[2,3], Andrew J. Saykin[4,5,6], R. Elizabeth Sanders[1], Laura E. Gibbons[7], Logan C. Dumitrescu[8,9], Michael L. Cuccaro[10,11], Timothy J. Hohman [8,9], Paul K. Crane[1]

[1] Department of Medicine, University of Washington, Seattle, WA, USA.
[2] Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, USA.
[3] VA Puget Sound Health Care System, GRECC, Seattle, WA, USA.
[4] Department of Radiology and Imaging Services, Indiana University School of Medicine, Indianapolis, IN, USA.
[5] Indiana Alzheimer Disease Center, Indiana University School of Medicine, Indianapolis, IN, USA.
[6] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.
[7] Department of General Internal Medicine, University of Washington, Seattle, WA, USA.
[8] Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Nashville, TN, USA.
[9] Vanderbilt Genetics Institute, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
[10] The John P. Hussman Institute for Human Genomics, University of Miami, Miami, FL, USA.
[11] Dr. John T. MacDonald Foundation, Department of Human Genetics, University of Miami, Miami, FL, USA.

Contents

Summary

This harmonization was recently published and provides additional details (1). Neuropsychological batteries differ across studies and cohorts, complicating the ability to merge cognitive data across cohorts. Co-calibrated composite scores generated using modern psychometric approaches permit direct comparison of study participants in different studies who were assessed with different test batteries. This document describes the co-calibration of memory, executive function, language, and visuospatial abilities composite scores across four cohorts: Alzheimer's Disease Neuroimaging Initiative (ADNI), Adult Changes in Thought (ACT), the Religious Orders Study and the Memory and Aging Project (ROS/MAP), and the National Alzheimer's Coordinating Center (NACC). The co-calibrated scores are standardized on the same metric, making the composite scores directly comparable even across studies that administered different tests. A schematic of the overall co-calibration workflow is presented in **Figure 1**.

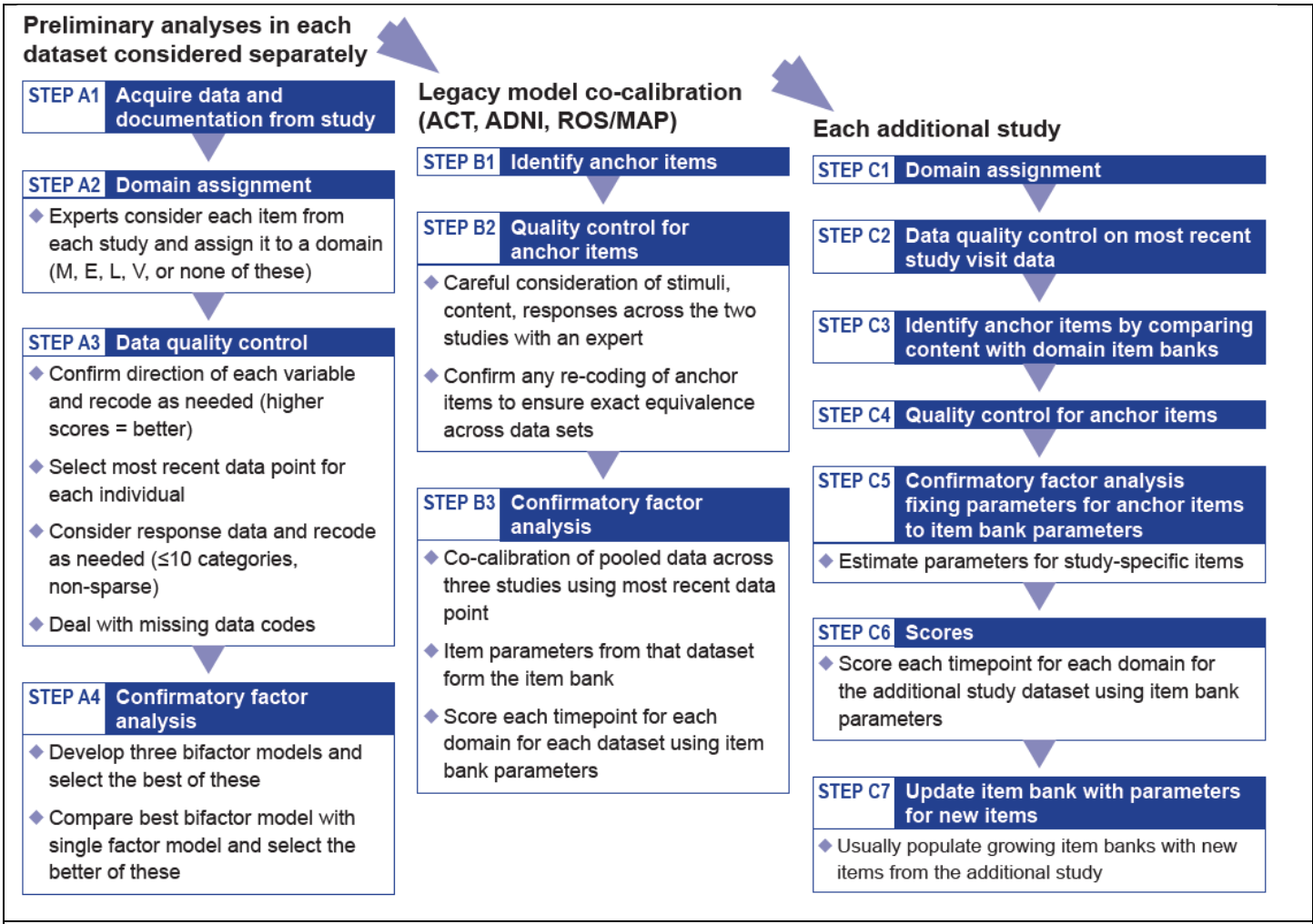


Figure 1. *Co-calibration workflow. Each of these steps is explained in more detail below.*

1. Preliminary analyses in each study included in the Legacy co-calibration model

"Legacy" here refers to the first set of studies we evaluated—ADNI, ACT, and ROS/MAP. These are three large, widely used studies with at least moderate sized neuropsychological batteries.

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, information on how responses are coded. Data dictionaries and protocol documents have proven to be useful. This step might take multiple iterations as we learn more about the data set.

STEP A2: Domain assignment. In each of the studies (Adult Changes in Thought [ACT], Alzheimer's Disease Neuroimaging Initiative [ADNI], the Religious Orders Study—Memory and Aging Project [ROS/MAP]), the expert panel (Dr. Trittschuh, Dr. Mez, and Dr. Saykin) assigned items from the neuropsychological battery to one of the four domains (memory, language, executive functioning, and visuospatial functioning). If applicable, the expert panel also assigned each of these items to sub-domains based on the cognitive processes involved in each task. Using study operational and administration manuals, as well as published results, we made note of differing versions and administration methods, so as to identify when the items were truly comparable. Not all items administered in these studies mapped to one of the four domains. Items that did not map to one of these domains were excluded in these analyses.

Note: On "Items"

We use granular data from each study. For a word list learning measure, one can imagine multiple ways of recording participant responses, including whether each specific word was recalled on each trial. Such data would be terrific for us. Many studies roll that sort of data into the number of words recalled on each trial. Those data would also be fine for us. And some studies total up all of the words recalled across all of the learning trials. Those data would be less valuable for us. In general, we work with parent studies to determine what data are digitally available, and work from there. Occasionally we find that data are not electronically available in a sufficiently granular form. In that instance we either seek resources for data entry or, if that proves impossible, we move to the next data set.

STEP A3: Data quality control: We obtained access to each cognitive datasets and Ms. Sanders, our data manager, performed initial quality control steps on the data. Before running psychometric models, we performed additional recoding of the data. Some items such as Trails A and B were reverse coded (i.e. where a higher value indicated lower performance). We checked each item to make sure lower values represent lower cognitive performance, and reverse coded as needed. The advantage of ensuring that higher scores always indicate better cognitive functioning is that negative loadings on factors stick out as possible sources of concern; better performance on an item should be associated with higher levels of the cognitive domain the item is an indicator of.

We selected the most recent visit for each participant when we had longitudinal data from a study. This choice optimizes the spread of cognitive abilities in the dataset. Some studies enrolled people known to be free of dementia (e.g. ACT and others), and others enrolled people with particular diagnoses with specific eligibility criteria (e.g. ADNI and others). A baseline dataset from studies with constrained enrollment (i.e. studies like ACT) would not include people with poorer cognitive functioning, so our calibrations from such datasets would be limited at the lower end of each domain. The last visit dataset optimizes the sample size (each person is included in the dataset) while optimizing the spread of cognitive abilities in each cohort (some people had intact cognitive functioning and some developed dementia by the last study visit).

We may need to revise this general approach (the "last visit" approach) in datasets where different batteries were included at different visits. For the legacy studies, the "last visit" approach resulted in a dataset with desirable properties, including a single observation per participant, and a maximal distribution of the underlying ability for each domain.

We considered the distribution of each item among participants with non-missing data and combined categories as needed. Our goals for combining categories were a) to avoid sparse categories (operationally defined as <5 responses for each study administering each item), b) to have a maximum of 10 categories, which is the maximum number of categories handled by *Mplus* v7.4(2), and c) to retain the full range of responses from each study, so we try to avoid collapsing categories at the highest and lowest levels of functioning (when there were at least 5 responses). Retaining variability at the tails at the expense of the center of the distribution minimizes floor and ceiling effects.

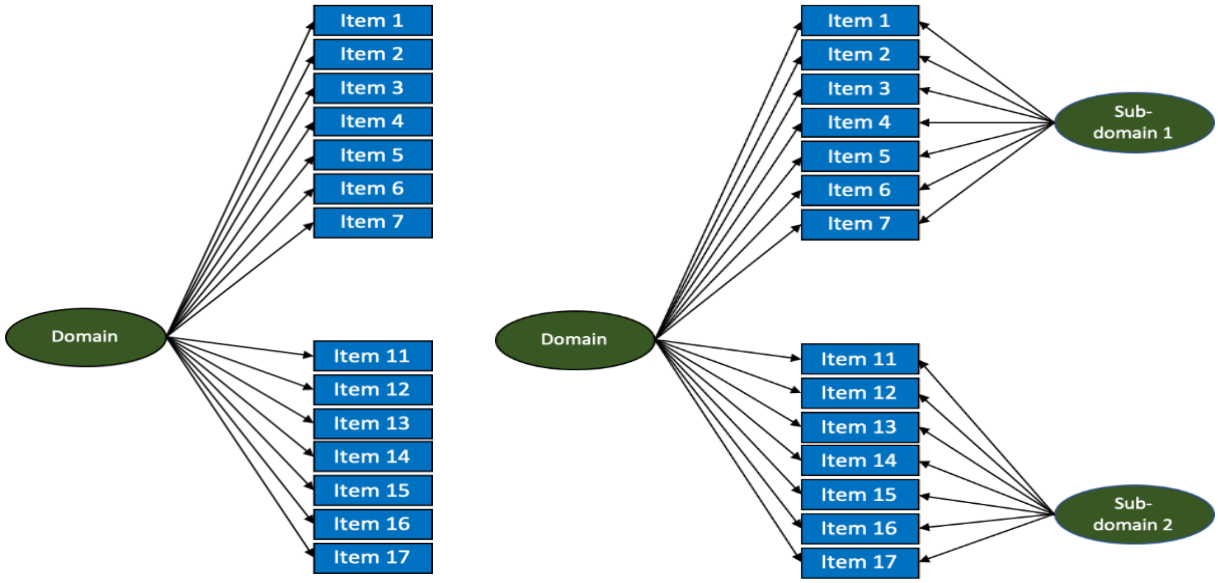
We treated each item as an ordinal indicator of the domain—the numerical value assigned to each category is irrelevant beyond its rank, e.g. calling the lowest category 3 points vs. 18 points makes no difference in how the item is treated or what the final score would be.

Missing data were a particular area of focus. Some studies had very little information beyond the fact of a missing data element. Other studies had specific distinct codes,

such as indicating participant refusal to complete an administered item vs. the interviewer ran out of time so the item was not administered. After careful consideration and sensitivity analyses, we ended up treating all types of missing data—regardless of codes available from the study—as if the item was not administered.

STEP A4: Confirmatory factor analyses: We used cognitive data from the last visit dataset in the legacy data sets to perform psychometric analyses to derive robust composite scores for each domain. We modeled each domain separately using confirmatory factor analysis with Mplus using a Robust Weighted Least Squares including terms for the mean and the variance (WLSMV) estimator. We ran two models: a.) a single factor model, with no residual structure; and b.) a bifactor model with hierarchical clustering-assigned sub-domains. A schematic representation of each of these models is provided in **Figure 2**. We consulted the expert panel on the sub-domain assignment of items to make sure these models made sense to our experts. We also considered theory-driven and methods effects-driven bifactor models. The data-driven bifactor model was superior to these two for all domains in all cases.

Figure 2. *Single factor (left) and bi-factor (right) models of 14 items from a single study*



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 depiction includes two subdomains (labeled "Sub-domain 1" and "Sub-domain 2"). These sub-domains 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 sub-domain 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 where the Domain in question was executive functioning), or a data-driven subset based on hierarchical clustering.

Our overall strategy in terms of single factor vs. bifactor modeling was that we would choose the single factor model if adding secondary factors did not markedly improve model fit and if adding secondary factors did not markedly impact any individual's score (see below).

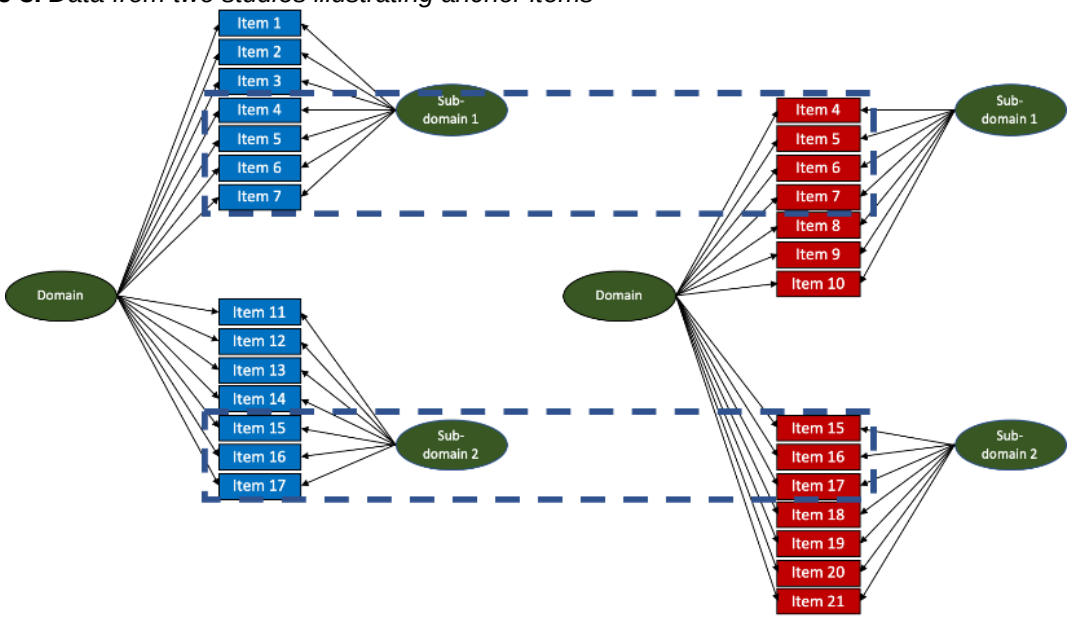
Our criteria for selecting the better model included evaluating fit statistics (see below) and concordance of model results with theory, such as positive loadings on secondary factors. The fit statistics we considered were the confirmatory fit index (CFI) where higher values indicate better fit; thresholds of 0.90 and 0.95 have been used in other settings as criteria for adequate or good fit(3, 4); the Tucker-Lewis Index (TLI), 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(3, 4).

When comparing the single factor model with the best bifactor model, we a) looked at 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 vs. 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 as the default level of tolerable measurement differences. 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.

2. Co-calibration of the domains across ACT, ADNI, AND ROS/MAP

STEP B1: Identification of anchor items: Co-calibration 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 confirmatory factor analysis models within each study. These 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**, below. We consulted the expert panel (Dr. Trittschuh) for anchor items selection review and confirmation.

Figure 3. *Data from two studies illustrating anchor items*



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 through 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). We will return to these studies in Figure 4 below.

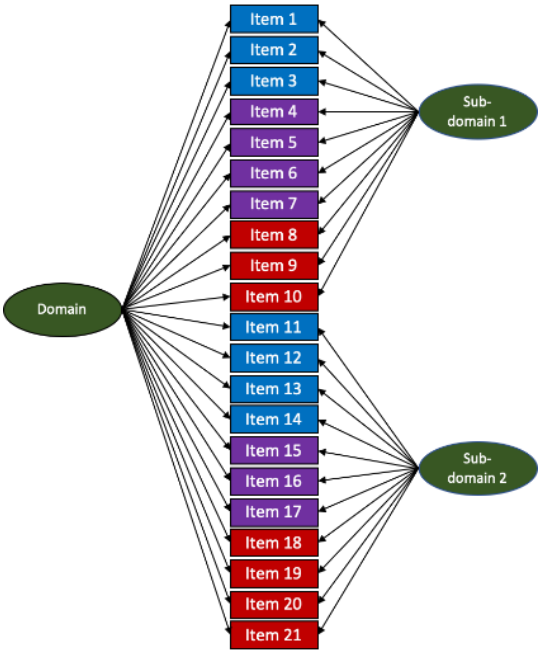
STEP B2: Quality control for anchor items: Anchor items were cleaned and recoded after considering item response data from all studies that administered the item, making sure that the range of responses to the anchor items was similar in each study. We carefully reviewed documentation from each study to ensure that the anchor item stimulus was precisely the same across studies, that the response options were precisely the same or could be re-coded to be the same 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 which study the person was enrolled in.

A note regarding response options—in many cases the stimulus is fairly open-ended, such as "can you please draw from memory the figure you copied a while ago," where the participant is handed a blank sheet of paper and a writing implement. The resulting drawing then gets scored based on how similar it was to the initial stimulus figure. The specific scoring applied to such a stimulus could vary across studies. One study could score such an item as binary (correct vs. incorrect), while another could apply points for various aspects of the drawing. We reviewed the scoring documentation from all studies administering any particular candidate anchor item to determine what "correct" meant in the first study, and how many aspects of the drawing would need to be present for a "correct" score in that study. Then we would map all scores from the second study that would have resulted in a "correct" score in the first study to a "correct" score, and all other scores from the second study to an "incorrect" score. In this way, the resulting score is invariant to which study the person is participating in, as each response would be consistently scored regardless of study.

STEP B3: Confirmatory factor analyses: We co-calibrated 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 mega-calibration model (see **Figure 4**).

One particularly tricky aspect of co-calibrating scores using bifactor models is how to handle secondary domains. Some anchor items had loadings on the primary domain (e.g. memory) and also 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 contains overlapping item(s) across studies along with study specific unique items, they were assigned to a common secondary domain in the mega-calibration model. While the CFA model with the WLSMV estimator produces fit statistics in M*plus*(2), the CFA model with the MLR estimator does not output fit statistics. We performed a number of sensitivity analyses to reassure ourselves that scores on the primary domain were minimally impacted by various ways of specifying the mean and variance on secondary domains. In the final models we selected, we specified a mean of 0 and a variance of 1 for each secondary domain factor, regardless of the number of studies that included items that loaded on that factor.

Figure 4. *Co-calibration of the red and blue studies.*



Co-calibration model for data from Study 1 and Study 2. Study 1 data includes blue and purple items, while Study 2 data include purple and red items. Purple items are anchors, which received extra attention and quality control (see above). This is referred to in this document as the "mega-calibration model".

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

Confirmatory factor analysis model co-calibration details

1. For all CFA models, we categorized items to ≤ 10 categories (the current limit for categorical variables in M*plus*). For co-calibration purposes, we had to re-categorize some of the items even though they already had ≤ 10 categories. This happened when some studies had more granular data (more categories) for anchor items compared to other studies. In these cases, after we estimated item parameters from the co-calibration model, we re-estimated parameters of the anchor item(s) in the most granular form in study or studies that had more response categories than the least common denominator categorization used for the anchoring analyses.

After using re-coded (less granular, least common denominator) items for co-calibration, we fixed all of the other items to their values from the co-calibration run and freely estimated parameters for re-coded anchors in their most granular form. This approach enabled us to obtain more precise scores in studies that incorporated more granular

scoring, while still using all items administered across studies to co-calibrate metrics across studies.

2. The base co-calibration model for each of the four domains was performed across ACT, ROS/MAP, and ADNI.

In these models,

a) The mean and variance for the primary factor were freely estimated.
b) If every item in a sub-domain in the new data had parameters available from the co-calibration model, we fixed those item parameters to their previously identified values, and allowed the mean and variance to be freely estimated in the new data.

If no item from a sub-domain had parameters available, then we freely estimated each of the sub-domain loadings, fixing the mean and variance of the subdomain factor to 0 and 1.

If there was a mix of previously specified and new items in a subdomain, we fixed the parameters for the previously specified items, and allowed the mean and variance of the factor and the loadings for new items to be freely estimated in the new data.

These details are necessary because of identifiability. In CFA models, either a loading and threshold need to be defined, or the mean and variance of a latent trait, for models to be identified. The first calibrations set latent trait means to 0 and variances to 1, obtaining item loadings and thresholds for all of the items. Then, when moving to other data from those original data sets, or to new data from new datasets, we treated the previously estimated item parameters as fixed, and freely estimated the mean and variance for the latent trait. This works whether one specifies only item parameters (loading and threshold(s)) from a single item or from multiple items; new items can be co-calibrated with previously calibrated anchor items using this approach.

Note: Detailed overview and all code snippets can be obtained from authors on request.

3. Scores for all time points for each data set

We used each study's item parameters from the mega-calibration model for a given domain (the item bank item parameters) to obtain scores for each person at each time point. Item parameters were forced into each of the legacy data sets separately, for each time point, with the mean and variance of the composite freely estimated, to derive factor scores for the primary factor (e.g. memory; labeled "Domain" in the figures) along with the corresponding standard error of measurement (SEM). We used all participants with relevant data to fit data for each domain, including people who may have been missing data entirely for some other domain.

For subsequent datasets in our pipeline such as the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS) etc. we ran CFA models at the pre-calibration step using data just from each study to determine the best fitting model if unique items were present which is not part of the item bank yet. We used data from the most recent study visit for calibration of data from that study. We compared items administered by the new study with items previously calibrated in our item bank. Items that had been administered in a study we had previously calibrated were treated as anchor items where we fixed their item parameters after thorough anchor item QC as described earlier. We freely estimated item parameters for items we had not previously encountered, meaning items not previously calibrated in our item bank. As before, once we had calibrated all the items from the new study using data from the most recent study visit, we used item parameters for all the items for a given domain to obtain scores with all of the longitudinal data to obtain domain scores and standard errors of measurement (SEMs) for each individual for each time point. In most cases we augmented our growing item banks with the item parameters from the non-anchor items in the subsequent dataset. If the distribution of available data was truncated, such as a study of younger adults or with very few cognitively impaired people, we did not augment the item bank.

Note: Studies such as NACC also has a large cohort below age 60 and our primary goals was to obtain scores for the age 60 and up population as best as we can. We are able to obtain scores for the younger (age < 60) sample separately too but with caveats. We used item parameters estimated from the 60 and up population on the younger sample while freely estimating the mean and variance of the resulting scores.

4. Standard errors and missing data in harmonized cognitive testing data

The bottom line: The harmonized domain-specific cognitive scores from each study were derived using modern psychometric modeling which produce scores and also the standard error of measurement (SEM) around those scores. After extensive discussion and consideration, our default is to include scores with SEM≦0.6 in common-use datasets. All scores including those with SEM >0.6 are available on request. Furthermore, the SEMs are also available for all scores; researchers wishing more precision than the default 0.6 may wish to choose a different SEM threshold for inclusion. The chosen value of 0.6 units for SEMs was data-driven, keeping a balance between length of neuropsychological battery, item discriminatory power, and item missingness.

Background and rationale: Measurement error is ubiquitous and frequently ignored. Cognitive tests also have measurement error. Modern psychometric modeling approaches provide tools to specifically address measurement error.

Intuitively, a brief battery of cognitive tests of a domain probably does not measure that domain as precisely as a long battery of cognitive tests. As a first approximation, there is likely a relationship between time of assessment for any particular domain and measurement precision.

Furthermore, modern psychometric approaches do not assume that measurement precision is identical for all research participants who receive the same battery of items. Again intuitively, if there were some high functioning people and some low functioning people, and a group of easy items, there would be little precision for the high functioning people (where all of the items were too easy) but more precision for the low functioning people. The density of item difficulty levels matched with the test taker's ability level are needed to determine the precision of measurement for each person even if the same items are administered. Modern psychometric approaches directly address item difficulty levels and person ability levels on the same metric, facilitating individualized estimation of measurement error / measurement precision.

Missing data also play a critical role in measurement precision in cognitive tests. Intuitively, imagine two people with identical ability levels, one of whom has complete data on all items of a domain and the other of whom has data on only half of the items. Ideally the estimated domain scores for the two individuals would be the same, but the precision should be more for the person with complete data.

Within studies, domains are measured with varying levels of intensity. In our experience with ADNI for example, memory is measured quite extensively (both the Rey and ADAS-Cog word lists, plus logical memory immediate and delayed, plus recognition, plus some MMSE items). At the other extreme, the measurement of visuospatial functioning is fairly limited in ADNI. Language and executive functioning are intermediate between those extremes in ADNI.

5. Neuropsychological items by domain for each study in Legacy model and fit statistics from CFA models

5A. Co-calibration of Memory

♦ **ACT:** Final model was the data driven bifactor model with CFI = 0.941, TLI = 0.936, and RMSEA = 0.057. The following items were included in the CFA analysis:

Table 1. Items and secondary structure for memory for the ACT study

Study	Variable	Description	Secondary Structure
ACT	mat_mem	DRS: Mattis Dementia Rating Scale Memory score	F1
ACT	w_in_c1	CERAD: Word list learning trial 1 total score	F1
ACT	w_in_c2	CERAD: Word list learning trial 2 total score	F1
ACT	w_in_c3	CERAD: Word list learning trial 3 total score	F1
ACT	w_rcl_c	CERAD: Word List Recall—correct	F1
ACT	w_rcg_t	CERAD: Word Recognition—total correct	F1
ACT	cp_re_ci	CERAD: Constructional Praxis Delay—circle	
ACT	cp_re_di	CERAD: Constructional Praxis Delay—diamond	
ACT	cp_re_re	CERAD: Constructional Praxis Delay—rectangles	
ACT	cp_re_cu	CERAD: Constructional Praxis Delay—cube	
ACT	w_lm_ima	WMS-R: Logical Mem I—immediate recall total story (AT)	F2
ACT	w_lm_imb	WMS-R: Logical Mem I—immediate recall total story (RM)	F3
ACT	w_lm_dea	WMS-R: Logical Mem II—delayed recall total story (AT)	F2
ACT	w_lm_deb	WMS-R: Logical Mem II—delayed recall total story (RM)	F3
ACT	w_vp_ine	WMS-R: Verbal Paired Associates I easy	F4
ACT	w_vp_inh	WMS-R: Verbal Paired Associates I hard	F5
ACT	w_vp_ree	WMS-R: Verbal Paired Associates II easy	F4
ACT	w_vp_reh	WMS-R: Verbal Paired Associates II hard	F5
ACT	rgs1	CASI: repeat words	
ACT	rc1a	CASI: Word recall—something to wear—1	F6
ACT	rc1b	CASI: Word recall—a color—1	F6
ACT	rc1c	CASI: Word recall—personal quality—1	F6
ACT	yr	CASI: What is today's date?—year	
ACT	mo	CASI: What is today's date?—month	F7
ACT	casi_dat	CASI: What is today's date?—day	F7
ACT	day	CASI: What day of week?	
ACT	casi_ssn	CASI: What season is it?	

ACT	spa	CASI: What state and city?	
ACT	spb	CASI: What is this place?	
ACT	rc2a	CASI: Word recall—something to wear—2	F6
ACT	rc2b	CASI: Word recall—a color—2	F6
ACT	rc2c	CASI: Word recall—personal quality—2	F6
ACT	rcobj	CASI: Recall of 5 objects	

♦ **ADNI:** Final model was a data driven bifactor model with CFI = 0.981, TLI = 0.979, and RMSEA = 0.088 for individuals administered RAVLT version A and CFI = 0.991, TLI = 0.990, and RMSEA = 0.073 for individuals administered RAVLT version B. The following items were included in the CFA analysis:

Table 2. Items and secondary structure for memory for the ADNI study

Study	Variable	Description	Secondary Structure
ADNI	limmtotal	WMS-R: Logical Memory—Immediate Recall(AT)	F1
ADNI	ldeltotal	WMS-R: Logical Memory—Delayed Recall (AT)	F1
ADNI	avtot1*	Rey: AVLT Trial 1 Total	
ADNI	avtot2*	Rey: AVLT Trial 2 Total	F2
ADNI	avtot3*	Rey: AVLT Trial 3 Total	F2
ADNI	avtot4*	Rey: AVLT Trial 4 Total	F2
ADNI	avtot5*	Rey: AVLT Trial 5 Total	F2
ADNI	avtot6*	Rey: AVLT Trial 6 Total	F2
ADNI	avtotb*	Rey: AVLT List B Total	
ADNI	avdel30min*	Rey: AVLT 30 Minute Delay Total	F2
ADNI	avdeltot*	Rey: AVLT Recognition Score	
ADNI	q1score	ADAS-Cog: Word Recall—score	F3
ADNI	q4score	ADAS-Cog: Delayed Word Recall	F3
ADNI	q7score	ADAS-Cog: Orientation—score	
ADNI	q8score	ADAS-Cog: Word Recognition—score	
ADNI	mmdate	MMSE: What is today's date?	
ADNI	mmyear	MMSE: What is the year?	
ADNI	mmmonth	MMSE: What is the month?	
ADNI	mmday	MMSE: What day of the week is today?	
ADNI	mmseason	MMSE: What season is it?	
ADNI	mmhospit	MMSE: What is the name of this hospital (clinic, place)?	
ADNI	mmfloor	MMSE: What floor are we on?	
ADNI	mmcity	MMSE: What town or city are we in?	
ADNI	mmarea	MMSE: What county (district, borough, area) are we in?	
ADNI	mmstate	MMSE: What state are we in?	

ADNI	bft1	MMSE: Ball, flag, tree—immediate recall (collapsed)	
ADNI	bft2	MMSE: Ball, flag, tree—delayed recall (collapsed)	
ADNI	mocaregi	MoCA: registration, sum of two trials	
ADNI	delsum	MoCA: delayed recall of word list	

* MoCA (blue) items were only administered in ADNI GO/2/3 while orange items were in all ADNI waves (1/GO/2).
* ADNI administered two versions (different word lists) of RAVLT (avtot1–avdeltot) and three different versions of ADAS-Cog items (q*) across waves. We ran the model separately for the two versions. The ADAS-Cog versions were found to be equivalent while the RAVLT versions were not. For determining secondary factor structures and extracting model fit statistics, we considered all RAVLT versions to be equivalent. The different versions of RAVLT were taken into account in the final co-calibration phase.
* There were additional MoCA items, which were the same (theoretically) as corresponding items from the Mini-Mental State Examination (MMSE). We excluded MoCA items if those items were already asked as part of the neuropsychological battery.

♦ **ROS/MAP:** Final model was a data driven bifactor model with CFI = 0.986, TLI = 0.984, and RMSEA = 0.081. The following items were included in the CFA analysis:

Table 3. Items and secondary structure for memory for the ROS/MAP study

Study	Variable	Description	Secondary Structure
ROS/MAP	mmse30_item1	MMSE: What is the year?	
ROS/MAP	mmse30_item2	MMSE: What is the season of the year?	
ROS/MAP	mmse30_item3	MMSE: What is the date?	
ROS/MAP	mmse30_item4	MMSE: What is the day of the week?	
ROS/MAP	mmse30_item5	MMSE: What is the month?	
ROS/MAP	mmse30_item6	MMSE: What state are we in?	
ROS/MAP	mmse30_item7	MMSE: What county are we in?	
ROS/MAP	mmse30_item8	MMSE: What city are we in?	
ROS/MAP	mmse30_item9	MMSE: What room are we in?	
ROS/MAP	mmse30_item10	MMSE: What is the address (street number and name) of this place?	
ROS/MAP	atb1	MMSE: apple, table, penny (immediate recall)	
ROS/MAP	atb2	MMSE: apple, table, penny (delayed recall)	
ROS/MAP	cts_wli1	CERAD: Word list learning Trial 1 (10 items collapsed)	F1
ROS/MAP	cts_wli2	CERAD: Word list learning Trial 2 (10 items collapsed)	F1
ROS/MAP	cts_wli3	CERAD: Word list learning Trial 3 (10 items collapsed)	F1

ROS/MAP	cts_wliii	CERAD: Word list recognition	
ROS/MAP	cts_wlii	CERAD: Word list recall	F1
ROS/MAP	cts_ebmt	East Boston Memory Test—immediate recall	F2
ROS/MAP	cts_ebdr	East Boston Memory Test—delayed recall	F2
ROS/MAP	cts_story	WMS-R: Logical memory (AT)	F3
ROS/MAP	cts_delay	WMS-R: Tell me the story again (AT)	F3

Co-calibration of memory across ACT, ADNI, ROS/MAP

Table 4. Co-calibration of memory across ACT, ADNI, ROS/MAP

Study	Variable	Secondary structure	Comments
ACT, ADNI, ROS/MAP	mmyear		mmse30_item1 in ROS/MAP; yr in ACT; mmyear in ADNI
ACT, ADNI, ROS/MAP	mmseason		mmse30_item2 in ROS/MAP; casi_ssn in ACT; mmseason in ADNI
ACT, ADNI, ROS/MAP	mmdate	F7	mmse30_item3 in ROS/MAP; casi_dat in ACT; mmdate in ADNI
ACT, ADNI, ROS/MAP	mmday		mmse30_item4 in ROS/MAP; day in ACT; mmday in ADNI
ACT, ADNI, ROS/MAP	mmmonth	F7	mmse30_item5 in ROS/MAP; mo in ACT; mmmonth in ADNI
ACT, ADNI, ROS/MAP	mmctst		Collapsed (mmse30_item6 mmse30_item8) in ROS/MAP and (mmcity mmstate) in ADNI to create a single variable; spa in ACT
ACT, ADNI, ROS/MAP	limmtotal	F2	limmtotal in ADNI; w_lm_ima in ACT; cts_story in ROS/MAP
ACT, ADNI, ROS/MAP	ldeltotal	F2	ldeltotal in ADNI; w_lm_dea in ACT; cts_delay in ROS/MAP
ROS/MAP	mmse30_item7		
ROS/MAP	mmse30_item9		
ROS/MAP	mmse30_item10		
ROS/MAP	atb1		
ROS/MAP	cts_story	F11	
ROS/MAP	cts_wli1	F10	
ROS/MAP	cts_wli2	F10	
ROS/MAP	cts_wli3	F10	
ROS/MAP	cts_ebmt	F12	
ROS/MAP	cts_wliii		
ROS/MAP	atb2		
ROS/MAP	cts_wlii	F10	
ROS/MAP	cts_ebdr	F12	
ROS/MAP	cts_delay	F11	
ACT	mat_mem	F6	
ACT	w_in_c1	F6	
ACT	w_in_c2	F6	

ACT	w_in_c3	F6	
ACT	w_rcl_c	F6	
ACT	w_rcg_t	F6	
ACT	cp_re_ci		
ACT	cp_re_di		
ACT	cp_re_re		
ACT	cp_re_cu		
ACT	w_lm_imb	F3	
ACT	w_lm_deb	F3	
ACT	w_vp_ine	F4	
ACT	w_vp_inh	F5	
ACT	w_vp_ree	F4	
ACT	w_vp_reh	F5	
ACT	rgs1		
ACT	rc1a	F1	
ACT	rc1b	F1	
ACT	rc1c	F1	
ACT	spb		
ACT	rc2a	F1	
ACT	rc2b	F1	
ACT	rc2c	F1	
ACT	rcobj		
ADNI	avtot1		Each RAVLT item was split into two items to account for two versions of RAVLT used in ADNI at specific waves where both versions of the same item were loaded into the same secondary structure
ADNI	avtot2	F8	
ADNI	avtot3	F8	
ADNI	avtot4	F8	
ADNI	avtot5	F8	
ADNI	avtot6	F8	
ADNI	avtotb		
ADNI	avdel30min	F8	
ADNI	avdeltot		
ADNI	q1score	F9	
ADNI	q4score	F9	
ADNI	q7score		
ADNI	q8score		
ADNI	mmhospit		
ADNI	mmfloor		
ADNI	mmarea		
ADNI	bft1		
ADNI	bft2		

ADNI	mocaregi		
ADNI	delsum		

5B. Co-calibration of Executive Functioning

♦ **ACT:** Final model was a data driven bifactor model with CFI = 0.748, TLI = 0.667, and RMSEA = 0.081. The following items were included in the CFA analysis:

Table 5. Items and secondary structure for executive functioning for the ACT study

Study	Variable	Description	Secondary Structure
ACT	mat_attn	DRS: Mattis Dementia Rating Scale, Attention score	
ACT	mat_conc	DRS: Mattis Dementia Rating Scale, Concentration score	
ACT	mat_ip	DRS: Mattis Dementia Rating Scale, initiation / perseveration score	
ACT	tr_a_tm	Trails A time to complete	F1
ACT	tr_b_tm	Trails B time to complete	F1
ACT	clockdr	Clock drawing	
ACT	dbsum	CASI: repeat numbers backward (3 trials collapsed)	F2
ACT	subtra	CASI: Subtraction (3 trials collapsed)	F2
ACT	sim	CASI: similarities	
ACT	jgmt	CASI: judgement	

♦ **ADNI:** Final models was a data driven bifactor model in ADNI 1 (CFI = 0.993, TLI = 0.990, and RMSEA = 0.050) and ADNI GO/2/3 (CFI = 0.972, TLI = 0.967, and RMSEA = 0.045). The following items were included in the CFA analysis:

Table 6. Items and secondary structure for executive functioning for the ADNI study

Study	Variable	Description	Secondary Structure
ADNI	clockcirc	Approximately circular face	
ADNI	clocksym	Symmetry of number placement	
ADNI	clocknum	Correctness of numbers	
ADNI	clockhand	Presence of the two hands	
ADNI	clocktime	Presence of the two hands, set to ten after eleven	
ADNI	dspanbac	WAIS-R: Digit Span Backward Total Correct	F2
ADNI	traascor	Trails A Time to Complete	F1
ADNI	trabscor	Trails B Time to complete	F1
ADNI	digitscor	WAIS-R: Digit Symbol Total Correct	F1
ADNI	dspanfor	WAIS-R: Digit Span Forward Total Correct	F2
ADNI	q13score	ADAS-Cog: Number cancellation task	F1
ADNI	mmrworld	MMSE: Spell WORLD backwards	

ADNI	absmeas	MoCA: Abstraction: watch-ruler	
ADNI	abstran	MoCA: Abstraction: train-bicycle	
ADNI	trails	MoCA: Trails	
ADNI	digback	MoCA: Digits Backward	
ADNI	serial	MoCA: Serial 7 total	
ADNI	digfor	MoCA: Digits Forward	
ADNI	letters	MoCA: List of Letters/Tapping: # Errors	

*MoCA (blue) items were only administered in ADNI GO/2/3 while gray items were in all ADNI waves (1/GO/2).

- **ROS/MAP:** Final model was a theory driven methods-effects model with CFI = 0.986, TLI = 0.976, and RMSEA = 0.104. The following items were included in the CFA analysis:

Table 7. Items and secondary structure for memory for the ROS and MAP studies

Study	Variable	Description	Secondary structure
ROS/MAP	cts_pmat	Raven Progressive Matrices composite	
ROS/MAP	cts_doperf	Digit Ordering composite	
ROS/MAP	mmse30_item26_dlr	MMSE: Spell WORLD backwards	
ROS/MAP	cts_db	WAIS-R: Digit Span Backward Total Correct	F2
ROS/MAP	cts_sdmt	Symbol digit modality test (oral)	F1
ROS/MAP	cts_nccrtd	Number comparison	F1
ROS/MAP	cts_df	WAIS-R: Digit Span Forward Total Correct	F2

Table 8. Co-calibration of executive functioning across ACT, ADNI, ROS/MAP

Study	Variable	Description	Secondary structure
ACT, ADNI	traascor	Trails A time to complete	F1
ACT, ADNI	trabscor	Trails B time to complete	F1
ADNI, ROS/MAP	dspanfor	Digit Span Forward: Total Correct	F2
ADNI, ROS/MAP	dspanbac	Digit Span Backward: Total Correct	F2
ADNI, ROS/MAP	mmrworld	MMSE: Spell WORLD backwards	
ROS/MAP	cts_pmat	Raven Progressive Matrices composite	
ROS/MAP	cts_doperf	Digit Ordering composite	
ROS/MAP	cts_sdmt	Symbol digits modality (oral)	F3
ROS/MAP	cts_nccrtd	Number comparison	F3
ACT	mat_attn	Mattis Dementia Rating Scale	
ACT	mat_conc	Mattis Dementia Rating Scale	
ACT	mat_ip	Mattis Dementia Rating Scale	
ACT	clockdr	Clock	
ACT	dbsum	CASI repeat numbers backward (3 trials collapsed)	

ACT	subtra	CASI Subtraction (3 trials collapsed)	
ACT	sim	CASI similarities	
ACT	jgmt	CASI judgement	
ADNI	clockcirc	Approximately circular face	
ADNI	clocksym	Symmetry of number placement	
ADNI	clocknum	Correctness of numbers	
ADNI	clockhand	Presence of the two hands	
ADNI	clocktime	Presence of the two hands, set to 10 after 11	
ADNI	digitscor	WAIS-R Digit Symbol Total Correct	F1
ADNI	q13score	ADAS-Cog: Number cancellation task	F1
ADNI	absmeas	MoCA Abstraction: watch-ruler	
ADNI	abstran	MoCA Abstraction: train-bicycle	
ADNI	trails	MoCA Trails	
ADNI	digback	MoCA Digits Backward	
ADNI	serial	MoCA Serial 7 total	
ADNI	digfor	MoCA Digits Forward	
ADNI	letters	MoCA List of Letters/Tapping: # Errors	

5C. Language Co-calibration of Language

♦ **ACT:** Final model was a data driven bifactor model with CFI = 0.962, TLI = 0.952, and RMSEA = 0.045. The following items were included in the CFA analysis:

able 9. Items and secondary structure for language for the ACT study

Study	Variable	Description	Secondary Structure
ACT	bnt_adpr*	Boston Naming Test – 10-item version	F1
ACT	bnt_cer *	CERAD: Boston Naming Test – 15-item version	F1
ACT	v_flu_t	Category Fluency (Animals) - Total Correct	
ACT	animal	CASI: animals with 4 legs	
ACT	rpta	CASI: repeat phrase 1	
ACT	rptb	CASI: repeat phrase 2	
ACT	cas_read	CASI: read and follow a command	
ACT	cas_writ	CASI: write a sentence	
ACT	cmd	CASI: obey oral commands	
ACT	body	CASI: identify parts of body	
ACT	obja	CASI: identify objects—1	
ACT	objb	CASI: identify objects—2	

* ACT administers all 15 items from the CERAD version of the Boston Naming Test (bnt_cer) and another 8 distinct items from a long version of the Boston Naming Test (bnt_adpr).

♦ **ADNI:** Final models was a single factor model in ADNI 1 (CFI = 0.979, TLI = 0.973, and RMSEA = 0.080); ADNI GO/2 (CFI = 0.977, TLI = 0.973, and RMSEA = 0.048), and ADNI 3 (CFI = 0.953, TLI = 0.943, and RMSEA = 0.037). The following items were included in the CFA analysis:

Table 10. Items and secondary structure for language for the ADNI study

Study	Variable	Description	Secondary structure
ADNI	catanimsc	Category Fluency (Animals) —Total Correct	
ADNI	catvegesc*	Category Fluency (Vegetables) —Total Correct	
ADNI	bnttotal**	BNT: Boston Naming Test: Total Number Correct (1+3)	
ADNI	q2score	ADAS-Cog: Commands	
ADNI	q5score	ADAS-Cog: Naming	
ADNI	q6score	ADAS-Cog: Ideational Praxis—score	
ADNI	mmrepeat	MMSE: Repeat after me: no ifs, ands, or buts.	
ADNI	mmhand	MMSE: Takes paper in right hand	
ADNI	mmfold	MMSE: Folds paper in half	
ADNI	mmonflr	MMSE: Puts paper on floor	
ADNI	mmread	MMSE: Present the piece of paper which reads	
ADNI	mmwrite	MMSE: Write a sentence.	
ADNI	camel	MoCA: Camel naming	
ADNI	lion	MoCA: Lion naming	
ADNI	rhino	MoCA: Rhinoceros naming	
ADNI	repeat1	MoCA: Repeat Sentence	
ADNI	repeat2	MoCA: Repeat Sentence	
ADNI	ffluency	MoCA: Letter Fluency—F (total number of correct words)	

* Only in ADNI 1; ** Boston Naming Test excluded in ADNI 3; Blue items are MoCA items introduced in ADNI GO/2/3.

* MMSE items watch and pencil naming were dropped from the model because of sparseness in cells. They are extremely easy items and <1% gets it wrong.

♦ **ROS/MAP:** Final model was a single factor model with CFI = 0.968, TLI = 0.964, and RMSEA = 0.063. The following items were included in the CFA analysis:

Table 11. Items and secondary structure for language for the ROS and MAP studies

Study	Variable	Description	Secondary Structure
ROS/MAP	mmse30_item26_wor	MMSE: Spell WORLD forwards	
ROS/MAP	mmse30_item17	MMSE: [SHOW WRIST WATCH] What is this called?	
ROS/MAP	mmse30_item18	MMSE: [SHOW PENCIL] What is this called?	
ROS/MAP	mmse30_item19	MMSE: Repeat a phrase	
ROS/MAP	mmse30_item20	MMSE: Read the words on this card, then do what it says	

ROS/MAP	mmse_cmd	MMSE: Takes piece of paper, fold in half, place in lap (3 items collapsed)	
ROS/MAP	mmse30_item24	MMSE: Write any complete sentence	
ROS/MAP	cts_bname	CERAD: Boston Naming Test – 15-item version	
ROS/MAP	cts_clothing	Category Fluency (Clothing) - Total Correct	
ROS/MAP	cts_animals	Category Fluency (Animals) - Total Correct	
ROS/MAP	cts_fruits	Category Fluency (Fruits) - Total Correct	
ROS/MAP	idea_item1	Complex ideation: Will a board sink in water?	
ROS/MAP	idea_item2	Complex ideation: Will a stone sink in water?	
ROS/MAP	idea_item3	Complex ideation: Is a hammer good for cutting wood?	
ROS/MAP	idea_item4	Complex ideation: Can you use a hammer to pound nails?	
ROS/MAP	idea_item5	Complex ideation: Do two pounds of flour weigh more than one?	
ROS/MAP	idea_item6	Complex ideation: Is one pound of flour heavier than two?	
ROS/MAP	idea_item7	Complex ideation: Will water go through a good pair of rubber boots?	
ROS/MAP	idea_item8	Complex ideation: Will a good pair of rubber boots keep water out?	

MMSE items watch and pencil naming were dropped from the legacy model since a) in ADNI those items were dropped because of sparseness in cells b) in ROSMAP, those items had very high standardized loadings (>0.98).

Table 12. Co-calibration of language across ACT, ADNI, ROS/MAP

Study	Variable	Description	Secondary structure
ACT, ADNI, ROS/MAP	read	MMSE: Read the words on this card, then do it	
ACT, ADNI, ROS/MAP	mmse_cmd	MMSE: Paper, fold, place on floor combined	
ACT, ADNI, ROS/MAP	catanim	Category Fluency (Animals)—Total Correct	
ACT, ROS/MAP	bnt_name	CERAD: Boston Naming Test – 15-item version	
ADNI, ROS/MAP	repeat	MMSE: Repeat a phrase	
ADNI, ROS/MAP	write	MMSE: Write any complete sentence	
ROS/MAP	mmse30_item26_wor	MMSE: Spell WORLD forwards	
ROS/MAP	cts_clothing	Category Fluency (Clothing) - Total Correct	
ROS/MAP	cts_fruits	Category Fluency (Fruits) - Total Correct	
ROS/MAP	idea_item1	Complex ideation: Will a board sink in water?	
ROS/MAP	idea_item2	Complex ideation: Will a stone sink in water?	
ROS/MAP	idea_item3	Complex ideation: Is a hammer good for cutting wood?	
ROS/MAP	idea_item4	Complex ideation: Can you use a hammer to pound nails?	
ROS/MAP	idea_item5	Complex ideation: Do two pounds of flour weigh more than one?	
ROS/MAP	idea_item6	Complex ideation: Is one pound of flour heavier than two?	

ROS/MAP	idea_item7	Complex ideation: Will water go through a good pair of rubber boots?	
ROS/MAP	idea_item8	Complex ideation: Will a good pair of rubber boots keep water out?	
ACT	bnt_adpr	Boston Naming Test – 10-item version	
ACT	animal	CASI: animals with 4 legs	
ACT	rpta	CASI: repeat phrase 1	
ACT	rptb	CASI:repeat phrase 2	
ACT	cas_writ	CASI: write something	
ACT	body	CASI: identify parts of body	
ACT	obja	CASI: identify objects—1	
ACT	objb	CASI: identify objects—2	
ADNI	catvegesc	Category Fluency (Vegetables) —Total Correct	
ADNI	bnttotal	Boston Naming Test: Total Number Correct (1+3)	
ADNI	q2score	ADAS-Cog Commands	
ADNI	q5score	ADAS-Cog Naming	
ADNI	q6score	ADAS-Cog: Ideational Praxis—score	
ADNI	camel	MoCA: Camel naming	
ADNI	lion	MoCA: Lion naming	
ADNI	rhino	MoCA: Rhinoceros naming	
ADNI	repeat1	MoCA: Repeat Sentence	
ADNI	repeat2	MoCA: Repeat Sentence	
ADNI	ffluency	MoCA: Letter Fluency—F (total number of correct words)	

5D. Co-calibration of Visuospatial functioning

♦ **ACT:** Final model was a single factor model with CFI = 1.000, TLI = 1.000, and RMSEA = 0.000. The following items were included in the CFA analysis:

Table 13. Items and secondary structure for visuospatial functioning for the ACT study

Study	Variable	Description	Secondary Structure
ACT	mat_cons	DRS: Mattis Dementia Rating Scale—constructional praxis score	
ACT	cp_in_ci	CERAD: Constructional Praxis—circle	
ACT	cp_in_di	CERAD: Constructional Praxis—diamond	
ACT	cp_in_re	CERAD: Constructional Praxis—rectangles	
ACT	cp_in_cu	CERAD: Constructional Praxis—cube	
ACT	draw	CASI: Copy interlocking pentagons	

♦ **ADNI:** Final model was a single factor model in ADNI 1/GO/2/3 with CFI = 0.988, TLI = 0.981, and RMSEA = 0.043. The following items were included in the CFA analysis:

Table 14. Items and secondary structure for visuospatial functioning for the ADNI study

Study	Variable	Description	Secondary Structure
ADNI	copycirc	Clock copy: Approximately circular face	
ADNI	copysym	Clock copy: Symmetry of number placement	
ADNI	copynum	Clock copy: Correctness of numbers	
ADNI	copytime	Clock copy: Presence of the two hands, set to ten after eleven	
ADNI	q3score	ADAS-Cog: Constructional Praxis—score	
ADNI	mmdraw	MMSE: Copy interlocking pentagons	

* Clock copy (copyhand) item was dropped from the model because of sparseness in cell. Almost all individuals got it correct.

♦ **ROS/MAP:** Final model was a single factor model with CFI = 0.940, TLI = 0.931, and RMSEA = 0.044. The following items were included in the CFA analysis:

Table 15. Items and secondary structure for visuospatial functioning for the ROS and MAP studies

Study	Variable	Description	Secondary Structure
ROS/MAP	mmse30_item25	MMSE: Copy interlocking pentagons	
ROS/MAP	lopair_item1-lopair_item15	JLO: Line orientation items (15 items)	

Co-calibration of visuospatial ability across ACT, ADNI, ROS/MAP

Table 16: Co-calibration of visuospatial functioning across ACT, ADNI, and ROS/MAP

Study	Variable	Description	Secondary Structure
ACT, ADNI, ROS/MAP	mmdraw	MMSE: Copy interlocking pentagons	
ROS/MAP	lopair_item1-lopair_item15	JLO: Line orientation items (15 items)	
ACT	mat_cons	DRS: Mattis Dementia Rating Scale—constructional praxis score	
ACT	cp_in_ci	CERAD: Constructional Praxis—circle	
ACT	cp_in_di	CERAD: Constructional Praxis—diamond	
ACT	cp_in_re	CERAD: Constructional Praxis—rectangles	
ACT	cp_in_cu	CERAD: Constructional Praxis—cube	
ADNI	copycirc	Clock copy: Approximately circular face	
ADNI	copysym	Clock copy: Symmetry of number placement	
ADNI	copynum	Clock copy: Correctness of numbers	
ADNI	copytime	Clock copy: Presence of the two hands, set to ten after eleven	
ADNI	q3score	ADAS-Cog: Constructional Praxis—score	

6. Addition of National Alzheimer's Coordinating Center (NACC) to the pipeline.

We dropped all telephone visits. NACC UDS sample is divided into two major subsets, UDS1/2 and UDS 3. UDS 1 and 2 have identical items. As detailed above (last paragraph of Text 2; Step B3), we used those co-calibrated item parameters for anchor items while freely estimating unique items administered only to NACC participants. The item parameter estimation step was performed separately for the UDS 1/2 and UDS 3 samples for individuals with age >= 60.

Once all item parameters were estimated, scores for a given domain along with standard error of measurement (SEM) were estimated for the combined UDS 1/2 and UDS 3 dataset for individuals with age >= 60. Following that, domain-level scores were generated for individuals with age < 60 using the same item parameters as above.

Note 1: NACC UDS battery did not have enough items for us to co-calibrate the visuospatial domain.

Note 2: These NACC domain scores should be treated as intermediate and not final. Each of the Alzheimer's Disease Research Centers (ADRCs) and Alzheimer's Disease Centers (ADCs) administer a longer neuropsychological battery than what is reported as part of UDS. We are in the process of obtaining those granular level data and using those to generate a more precise score for each domain, as well as generate score for visuospatial functioning if enough items for that domain has been administered.

- **MEMORY:** ADNI and ROS/MAP recorded granular item-level data for all tests. For Mini-Mental State Examination (MMSE), NACC UDS 1/2 reported subscales for time and place. The time subscale is sum of five MMSE items (month, date, year, say, and season) while the place subscale is the sum of another five MMSE items (hospital, floor, city, area, and state). For the MoCA delayed recall test (no cue) (face, velvet, church, daisy, red), NACC UDS 3 recorded total scores (0-5) in a different way than in ADNI (0-15). We were able to construct these two MMSE subscales in ADNI and ROS/MAP and MoCA recall test in ADNI by using granular data. Additionally, we were able to use MoCA orientation items – city and state as anchors since we were able to use those granular items from ROS/MAP. We re-ran the legacy model by fixing all other items to their initial parameters and freely estimating these newly derived items so that these can be treated as anchors.

  UDS 1/2: We chose a bifactor model shown below with residual correlation between the two logical memory items. All items were part of the item bank.

  UDS 3: We chose a bifactor model shown below with CFI = 0.998, TLI = 0.997, and RMSEA= 0.036.

Table 17: Items and secondary structure for memory in NACC. Anchor items previously encountered in the item bank are shown in orange and novel items are shown in green.

Study	Variable	description	Secondary structure
ACT, ADNI, NACC (UDS 1/2/3)	logimem	Logical Memory Immediate A: Story units recalled (Story A ǂ)	F1
ACT, ADNI, NACC (UDS 1/2/3)	memunits	Logical Memory Delayed A: Story units recalled (Story A ǂ)	F1
ADNI, ROS/MAP, NACC (UDS 1/2)	mmseorda	MMSE Orientation subscale score – time	
ADNI, ROS/MAP, NACC (UDS 1/2)	mmseorlo	MMSE Orientation subscale score – place	
ACT, ADNI, ROS/MAP, NACC (UDS 3)	mocaordt	MoCA orientation - date	
ACT, ADNI, ROS/MAP, NACC (UDS 3)	mocaordy	MoCA orientation - day	
ACT, ADNI, ROS/MAP, NACC (UDS 3)	mocaormo	MoCA orientation - month	
ACT, ADNI, ROS/MAP, NACC (UDS 3)	mocaryr	MoCA orientation – year	
ROS/MAP, NACC (UDS 3)	mocaorct	**MoCA orientation - city	
ROS/MAP, NACC (UDS 3)	mocaorpl	MoCA orientation - place	
ADNI, NACC (UDS 3)	mocarecn	MoCA delayed recall - no cue	
ADNI, NACC (UDS 3)	mocaregi	MoCA memory: registration (2 trials)	
NACC (UDS 3)	craftdvr	Craft story 21 recall (delayed): paraphrase	F2
NACC (UDS 3)	craftvrs	Craft story 21 recall (delayed): verbatim	F2
NACC (UDS 3)	udsbenrs	Benson complex figure recognition	
NACC (UDS 3)	udsbentd	Benson complex figure copy, delayed	

* MMSE: Mini-Mental State Examination; ** MoCA: Montreal Cognitive Assessment; ǂ Story A: Anna Thompson

♦ **LANGUAGE:** ADNI recorded granular item-level data for all tests. For the Montreal Cognitive Assessment (MoCA) test, NACC UDS 3 collapsed items for naming (lion, rhino, and camel) and for two repetition trials. We were able to construct these two MoCA items in ADNI by using granular data. We re-ran the legacy model by fixing all other items to their initial parameters and freely estimating these newly derived items so that these can be treated as anchors.

UDS 1/2: We used a bifactor model shown below. All items were part of the item bank.

UDS 3: We chose a bifactor model shown below with CFI = 0.990, TLI = 0.984, and RMSEA= 0.084.

Table 18: Items and secondary structure for language in NACC. Anchor items previously encountered in the item bank are shown in orange and novel items are shown in green.

Study	Variable	Description	Secondary structure
ACT, ADNI, ROS/MAP, NACC (UDS 1/2/3)	catanimsc	Category Fluency (Animals) – Total Correct	
ADNI, NACC (UDS 1/2/3)	catvegesc	Category Fluency (Vegetables) – Total Correct	
ADNI, NACC (UDS 1/2)	bnttotal	Boston Naming Test: Total number correct (1+3)	
ADNI, NACC (UDS 3)	mocanami	MoCA: language – naming (lion, camel, rhino)	
ADNI, NACC (UDS 3)	ffluency	*MAE: Letter Fluency – F (total number of correct words)	F1
ADNI, NACC (UDS 3)	mocarepe	**MoCA: language – repetition	
NACC (UDS 3)	mintots	Multilingual naming test (MINT) – total score	
NACC (UDS 3)	udsverlc	MAE: Letter Fluency – L (total number of correct words)	F1

* MAE: Multilingual Aphasia Examination; ** MoCA: Montreal Cognitive Assessment

References

1. Mukherjee S, Choi S-E, Lee M, Scollard P, Trittschuh EH, Mez J, Saykin AJ, Gibbons LE, Sanders RE, Zaman AF, Teylan MA, Kukull WA, Barnes LL, Bennet DA, Lacroix A, Larson EB, Cuccaro M, Mercado S, Dumitrescu L, Hohman TJ, Investigators from ACT, ADNI**, ROS, MAP, MARS, NACC, and Crane, PK, (2022): Cognitive domain harmonization and co-calibration in studies of older adults. *Neuropsychology*. DOI: 10.1037/neu0000835. PMC Pending.

2. Muthén LK, Muthén BO. Mplus: statistical analysis with latent variables. 5.1 ed. Los Angeles, CA: Muthén & Muthén; 1998-2007.

3. Hu L-t, Bentler PM. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. Structural Equation Modeling. 1999;6(1):1-55.

4. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl 1):S22-31.

Version Information

This is the third version of this document.

Dataset Information

This methods document applies to the following dataset(s) available from the ADNI repository:

Dataset Name	Date Submitted
ADSP Phenotype Harmonization Consortium (PHC) - Composite Cognitive Scores	5 October 2022

About the Authors

This document was prepared by Dr. Mukherjee and Dr. Crane from the University of Washington, along with Dr. Hohman and Dr. Dumitrescu from Vanderbilt University Medical Center. For more information, please contact Dr. Hohman by email at timothy.j.hohman@vumc.org.